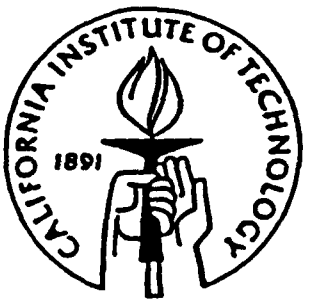
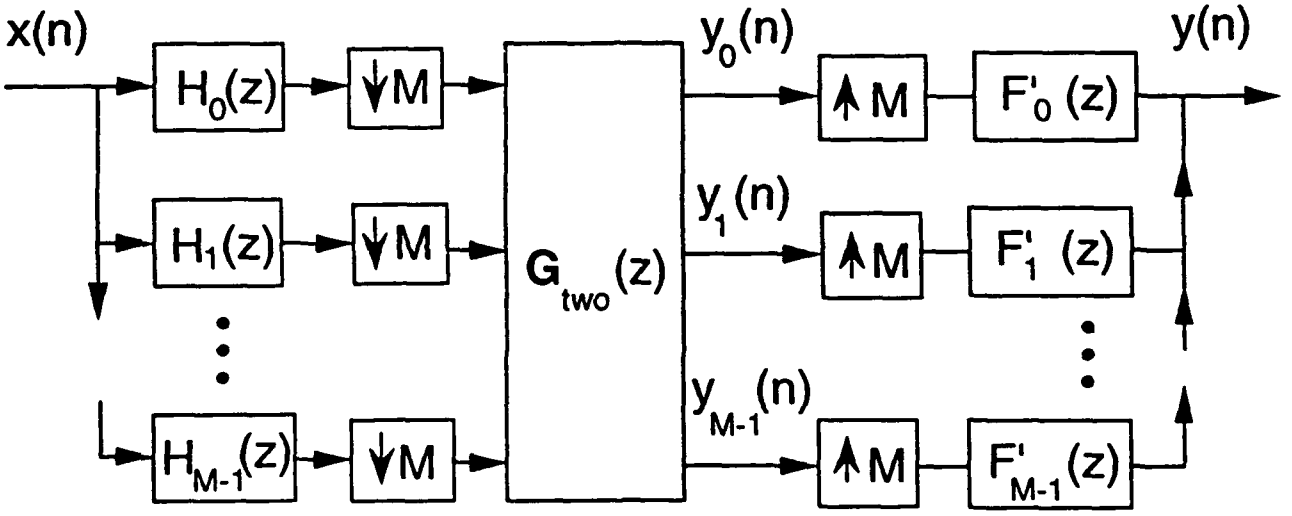


March 1993

ONE- AND TWO-LEVEL FILTER BANK CONVOLVERS

See-May Phoong and P. P. Vaidyanathan

Department of Electrical Engineering



A unification of two-level filter bank convolver with digital block filtering.

California Institute of Technology
Pasadena, CA 91125

ONE- AND TWO-LEVEL FILTER-BANK CONVOLVERS‡

See-May Phoong, Student Member, IEEE and P. P. Vaidyanathan, Fellow, IEEE

Department of Electrical Engineering, 116-81

California Institute of Technology

Pasadena, CA 91125, USA

Abstract. In a recent paper, it was shown in detail that in the case of orthonormal and biorthonormal filter banks we can convolve two signals by directly convolving the subband signals and combining the results. In this paper, we further generalize the result. We also derive the *statistical* coding gain for the generalized subband convolver. As an application, we derive a novel low sensitivity structure for FIR filters from the convolution theorem. We define and derive a *deterministic* coding gain of the subband convolver over direct convolution for a fixed wordlength implementation. This gain serves as a figure of merit for the low sensitivity structure. Several numerical examples are included to demonstrate the usefulness of these ideas. By using the generalized polyphase representation, we show that the subband convolvers, linear periodically time varying systems, and digital block filtering can be viewed in a unified manner. Furthermore, the scheme called IFIR filtering is shown to be a special case of the convolver.

EDICS number: SP 2.4.5

‡ This work was supported by the Office of Naval Research Grant N00014-93-1-0231, and funds from Tektronix, Inc.

1. INTRODUCTION

1.1. Main results of this paper and previous work

Convolution plays a central role in digital signal processing. Many well-known algorithms are proposed to reduce the computational complexity of convolution [1]. In this paper, our aim is not to find an algorithm that is faster than existing fast algorithms. Our goal is to find a more accurate way to compute the convolution when the convolution is implemented with finite precision. For this we use filter bank techniques. Consider the filter bank in Fig. 1.1(a), where $H_k(z)$ are the analysis filters and $F_k(z)$ are the synthesis filters. This multirate system has been studied by a number for researchers [2]-[8]. The analysis bank $\{H_k(z)\}$ splits the signal $x(n)$ into the subband signals $x_k(n)$.

In a recent paper [2], it was shown that if the systems in Fig. 1.1 are perfect reconstruction systems (i. e. $x(n) = \widehat{x}(n)$ and $g(n-i) = \widehat{g}(n-i)$), we can obtain the convolution of $x(n)$ and $g(n)$ by simply convolving $x_k(n)$ and $g_k^{(i)}(n)$ and adding the results. No *cross*-convolution between the subband signals is involved. When the computation is done with finite percision, it was also shown in [2] how the energy distribution in the subbands of $x(n)$ and $g(n)$ can be exploited to obtain a more accurate (compared to direct convolution) result. Optimal bit allocation and coding gain over the direct convolution were derived for both the cases of uniform and nonuniform decimated systems. In this paper, we further generalize the subband convolution theorem and show that no cross-convolution between the subband signals is involved in the generalized subband convolution theorem. Then we derive the coding gain for this generalized subband convolver. We will also show that the coding gain for the generalized convolver is always greater than that derived in [2], provided that the filter banks are orthonormal. We will refer to the convolution theorem derived in [2] as *one-level* filter bank (FB) convolution theorem and the generalized theorem in this paper as *two-level* FB convolution theorem. The reason for these names will be made clear in Section 2.

In [2], only the quantization in the subbands of $x(n)$ was considered. In this paper, we will address the case when the subband signals of $g(n)$ are quantized. In this case, we quantize the filter coefficients $g_k^{(i)}(n)$ in the subbands based on the input signal variance, and maximum amplitude of the filter. In the process of quantization, the filter coefficients are treated as *deterministic* parameters instead of random variables as done in [9]. Thus overflow of subband coefficients is completely avoided. We will derive the optimal bit allocation and the deterministic coding gain formulas. The derivation leads to a novel low sensitivity structure for FIR filters. The new structure is particularly attractive when the filter $g(n)$ is frequency selective and has a long impulse response, or it has some special time-frequency relation, e.g. the matched filtering of a chirp signal in radar application [10].

In this paper, we also explore the relationship between the convolver and the digital block filtering [11], [12] and [13]. We show that both the one-level and two-level FB convolvers are generalizations of the conventional block filtering. The subband convolvers have both the advantages of coding gain and parallelism. Adaptive filtering in subbands has been introduced with the goals of both reducing the computational

complexity and improving the convergence speed of the algorithm. In [14], it was observed that the overall performance of subband adaptive filtering is acceptable though there is some degradation in the convergence performance. In the view of generalized block filtering, the structure used in [14] can be regarded as a simplified version of the two-level FB convolver introduced in this paper. Thus it is possible to improve the performance of the subband adaptive filter by using the two-level FB convolver.

In the light of block filtering, we generalize the subband convolution technique to implement a linear periodically time varying (LPTV) filter. By using the generalized polyphase representation, we show that interpolated finite impulse response (IFIR) filter [15] is a special case of the subband convolver.

The filter bank techniques have been used in [16] to implement FIR and IIR filters. A different subband convolution theorem which leads to computational saving is derived in [16]. The subband convolution theorem proposed is applied to the digital pulse compression in radar application by Steffen in [10] and the convolution using DFT filter bank is discussed in detail. The subband convolution theorems discussed in this paper and in [2] differ from that derived in [10] and [16] in the sense that the convolution is *'perfect'* regardless of filter responses of the biorthonormal filter bank. Moreover it works for the nonuniform and maximally decimated cases.

1.2. Outline of the paper

Our presentation will go as follows:

1. In Section 2, we will generalize the subband convolution theorem. This is the two-level FB convolution theorem. A pictorial proof of the theorem is provided in Section 2.3 to give a clearer insight into what is going on in the convolution theorem.
2. In Section 3, we consider the quantization of the input signal $x(n)$. The optimal bit allocation and coding gain for the two-level FB convolver are presented.
3. A low sensitivity structure is derived in Section 4, first using the one-level FB convolver and then the two-level FB convolver. The optimal bit allocation and deterministic coding gain formulas for both the cases will be derived.
4. Several numerical examples are included in Section 5 to demonstrate the usefulness of the low sensitivity structures. From the examples, we will see that the performance of the two-level FB convolver is better than that of the one-level FB convolver. The coding gain of the convolvers is also shown.
5. We will discuss the relationship among conventional block filtering, and one-level and two-level FB convolvers in Section 6.1. In the presence of quantizers in the subbands of $g(n)$, we will show in Section 6 that the linear time invariant (LTI) filter is effectively replaced by a LPTV filter in the low sensitivity implementation. We will analyze the effect of this.
6. In Section 7, we will consider the application of the subband convolution theorem to infinite impulse response (IIR) filters.

7. In the last section, we will relate the IFIR filter to the subband convolver.

1.3. Notations and preliminaries

Notations: Capital boldfaced letters and lowercase boldfaced letters are used to denote matrices and vectors respectively. The (k, i)th element of a matrix $\mathbf{E}$ is denoted by $[\mathbf{E}]_{ki}$. The superscript * denotes complex conjugate and $\dagger$ denotes conjugate followed by tranposition. The z-transform of $v(n)$ is represented by $V(z)$. The notations $\big(V(z)\big)_{\downarrow M}$ and $\big(V(z)\big)_{\uparrow M}$ denote the M-fold decimated and M-fold expanded versions of the signal $v(n)$ respectively. The convolution of $x(n)$ and $g(n)$ is denoted by $x(n) * g(n)$.

Quantizers: Consider Fig. 1.1(a), where a general M-channel nonuniform decimated filter bank is shown. The boxes labelled Q_i are the quantizers. By b bit quantizer, we mean that the output signal of the quantizer is represented by b bits plus a sign bit. In this paper, the weight on the most significant bit is fixed for a fixed quantizer.

The decimators and expanders [17]: The boxes with $\downarrow n_k$ denote the n_k-fold decimators and the boxes with $\uparrow n_k$ denote the n_k-fold expanders. Their operations can be mathematically described respectively by the following two equations:

$$\Big(V(z)\Big)_{\downarrow n_k} = \frac{1}{n_k}\sum_{i=0}^{n_k-1} V\big(z^{1/n_k}W_{n_k}^i\big), \quad \Big(V(z)\Big)_{\uparrow n_k} = V\big(z^{n_k}\big), \tag{1.1}$$

where $W_{n_k} = e^{-j2\pi/n_k}$. The subscript n_k on W will be omitted whenever it is clear in the discussion.

Maximal decimation: An M-channel nonuniform multirate system is said to be *maximally* decimated if $\sum_{i=0}^{M-1}\frac{1}{n_k} = 1$. In the uniform case where all n_k are equal, this translates to $n_k = M$ for all k.

Conventional polyphase representations [3], [18], [17]: Consider a set of filters $H_k(z)$, $k = 0, 1, \ldots M-1$. They can be uniquely written in terms of their M polyphase components as $H_k(z) = \sum_{l=0}^{M-1} z^{-l}E_{kl}(z^M)$. This is known as Type 1 polyphase representation and $E_{kl}(z)$ is called the l-th polyphase component of $H_k(z)$. The $M \times M$ matrix $\mathbf{E}(z)$, with its k-th row l-th column element $[\mathbf{E}(z)]_{kl} = E_{kl}(z)$, is called the Type 1 polyphase matrix of the filters $H_k(z)$. Similarly, $H_k(z)$ can be written in terms of their Type 2 polyphase components as $H_k(z) = \sum_{l=0}^{M-1} z^l R_{lk}(z^M)$. The Type 2 polyphase matrix $\mathbf{R}(z)$ of the filters $H_k(z)$ is defined as $[\mathbf{R}(z)]_{lk} = R_{lk}(z)$. These polyphase representations are proved to be valuable in both the theory and design of filter banks.

Generalized polyphase representations [6], [19]: Generalized polyphase (GPP) was introduced in [6] and used in [19] to enchance the coding gain of subband coding. Instead of expressing a signal $v(n)$ in terms of the functions $\{z^{-i}\}$ as in the conventional polyphase representation, we express $v(n)$ in terms of the functions $\{U_i(z)\}$ as follows:

$$V(z) = \sum_{i=0}^{M-1} V_i(z^M)U_i(z). \tag{1.2}$$

This representation is said to be a valid GPP representation if the functions $\{U_i(z)\}$ (called a "polyphase basis") satisfy the conditions [19]: (i) Every rational function $V(z)$ can be expressed as (1.2), where $V_i(z)$

are rational. (ii) $V(z)$ is FIR if and only if $V_i(z)$ are FIR. By defining $\mathbf{U}(z)$ to be the conventional polyphase matrix of the polyphase basis $\{U_i(z)\}$, these conditions were proved to be equivalent to $\det[\mathbf{U}(z)] = cz^k$, for $c \neq 0$ and integer k. We will call $V_i(z)$ the ith GPP component of the signal $v(n)$ with respect to the polyphase basis $\{U_i(z)\}$.

Orthonormality and biorthonormality [2], [3], [8], [20]: From Fig. 1.1(a) (ignore the quantizers in the subbands), the z-transform of the output of the filter bank is

$$\widehat{X}(z) = \sum_{k=0}^{M-1} X_k(z^{n_k}) F_k(z). \tag{1.3}$$

If $\widehat{x}(n) = x(n)$ for all $x(n)$, then the system is called a biorthonormal or perfect reconstruction filter bank. The biorthonormality of the filter bank translates to the following condition on the filters $H_k(z)$ and $F_k(z)$:

$$\Big[H_k(z)F_m(z)\Big]_{\downarrow n_{k,m}} = \delta(k-m), \tag{1.4}$$

where $n_{k,m} = \gcd(n_k, n_m)$. From (1.4), we can see that the perfect reconstruction property is preserved when we interchange the roles of $H_k(z)$ and $F_k(z)$. The set of filters $\{F_k(z)\}$ is said to be orthonormal if $\big(F_k(z)F_m^*(1/z^*)\big)_{\downarrow n_{k,m}} = \delta(k-m)$. In the case of uniform decimation system, the filter bank is said to be paraunitary if the polyphase matrix $\mathbf{R}(z)$ of $\{F_k(z)\}$ satisfies the condition $\mathbf{R}(z)\mathbf{R}^\dagger(1/z^*) = \mathbf{I}$, where $\mathbf{I}$ is identity matrix.

Remarks: For uniform biorthonormal system (all $n_k = M$) with FIR analysis and synthesis filters, (1.3) is a GPP representation of $X(z)$ with the polyphase basis $\{U_i(z)\}$ taken to be $\{F_i(z)\}$. The subband signals $X_k(z)$ can be regarded as the GPP components.

An M-channel delay chain: Consider Fig. 1.1(a). Let all $n_k = M$. Then the system is a uniformly decimated system. If the filters $H_k(z) = z^{-k}$ and $F_k(z) = z^k$, then the system is called an M-channel delay chain. Notice that a delay chain is trivially an orthonormal system with the identity matrix $\mathbf{I}$ as the polyphase matrix.

2. ONE- AND TWO-LEVEL FILTER-BANK CONVOLUTION THEOREM

2.1. Review of one-level FB convolution theorem

Consider the two maximally decimated filter banks as shown in Fig. 1.1 (ignore the quantizers in the discussion of this section). Assume that the system has perfect reconstruction. Then it was shown in [2] how we can convolve two signals $x(n)$ and $g(n)$ by directly convolving the subband signals $x_k(n)$ and $g_k^{(i)}(n)$ and adding the results. *No cross-coupling* between subbands is involved. More precisely, we have the following biorthonormal convolution theorem:

Theorem 2.1 [2]. ***One-level* filter bank (FB) convolver.** Consider Fig. 1.1. Assume that the system has perfect reconstruction. Define the integer $p_k = L/n_k$, where L denotes the lcm of the decimation ratios

$\{n_k\}$. Let $x_k(n)$ and $g_k^{(i)}(n)$ be the subband signals defined in Fig. 1.1(a) and (b) respectively. Then the ith polyphase component, $y_i(n)$ of $x(n) * g(n)$ can be written as

$$y_i(n) = \Big(x(n) * g(n-i)\Big)_{\downarrow L} = \sum_{k=0}^{M-1} \Big(x_k(n) * g_k^{(i)}(n)\Big)_{\downarrow p_k}. \tag{2.1}$$

◇

The advantage of the subband convolution is that we can compute the result more accurately when the convolution is implemented with finite precision. It was shown in [2] how we can quantize the subband signals $x_k(n)$, and reduce the quantization noise by optimally allocating the bits in the subbands. By exploiting the subband energy distribution, the optimal bit allocation scheme and the coding gain over direct convolution were derived in [2].

Comments on complexity. Notice that the subband convolution theorem holds even when the analysis and the synthesis filters are IIR filters. But if we consider computational cost, the FB convolver is useful only when $H_k(z)$ and $F_k(z)$ are FIR filters. Thus in this paper, we will consider subband convolvers with FIR analysis and synthesis filters only. Also note that the computation of $x_k(n)$ involves filtering. Since $g(n)$ is a fixed filter, the subband signals $g_k^{(i)}(n)$ can always be precomputed and stored. Thus the complexity of the subband convolution is approximately equal to that of direct convolution plus the cost of implementing an analysis bank, assuming that no fast algorithm for convolution is used. If the complexity of the filter bank is low (compared to the length of the sequences $x(n)$ and $g(n)$), then the computational cost of $x_k(n)$ is negligible compared to that of the convolution. In this case the complexity of subband convolution and that of direct convolution are approximately the same.

2.2. Two-level FB convolution theorem

Theorem 2.2. *Two-level* **FB convolver.** Consider Fig. 1.1. Let $\{H_k(z)\}$ and $\{F_k(z)\}$ be respectively the analysis and synthesis filters of an M-channel maximally decimated nonuniform filter bank with perfect reconstruction. Define the integer $p_k = L/n_k$, where L denotes the lcm of the decimation ratios $\{n_k\}$. Let $\{H'_k(z)\}$ and $\{F'_k(z)\}$ be respectively the analysis and synthesis filters of a "L-channel" uniform biorthonormal system. Let $x_k(n)$ and $g_k^{(i)}(n)$ be respectively the kth subband signals defined in Fig. 1.1(a) and Fig. 2.1. Then the ith GPP component $y_i(n)$ of $x(n) * g(n)$ with respect to the polyphase basis $\{F'_i(z),\ i = 0,\ 1,\ \ldots,\ L-1\}$ can be written as

$$y_i(n) = \sum_{k=0}^{M-1} \Big(x_k(n) * g_k^{(i)}(n)\Big)_{\downarrow p_k}. \tag{2.2}$$

◇

Proof. From the definition of $X_k(z)$ and $G_k^{(i)}(z)$ and using the biorthonormality (1.3) of the filters $\{H_k(z)\}$ and $\{F_k(z)\}$, we get

$$X(z) = \sum_{k=0}^{M-1} X_k(z^{n_k}) F_k(z), \quad \text{and} \quad G(z)H'_i(z) = \sum_{l=0}^{M-1} G_l^{(i)}(z^{n_l}) H_l(z). \tag{2.3}$$

Multiplying the above two equations and decimating both sides by L, we have

$$\left[X(z)G(z)H_i'(z)\right]_{\downarrow L} = \sum_{k=0}^{M-1}\sum_{l=0}^{M-1}\left[X_k(z^{n_k/n_{k,l}})G_l^{(i)}(z^{n_l/n_{k,l}})\right]_{\downarrow p_{k,l}}\left[F_k(z)H_l(z)\right]_{\downarrow L}$$
$$= \sum_{k=0}^{M-1}\left[X_k(z)G_k^{(i)}(z)\right]_{\downarrow p_k}, \tag{2.4}$$

where $n_{k,l} = \gcd(n_k, n_l)$ and the integer $p_{k,l} = L/n_{k,l}$. The second equality follows from (1.4) and the fact that $\left[V(z)\right]_{\downarrow L} = \left[\left(V(z)\right)_{\downarrow n_{k,l}}\right]_{\downarrow p_{k,l}}$. Applying the biorthonormality of the filters $\{H_i'(z)\}$ and $\{F_i(z)\}$ the left hand side of (2.4) is by definition the ith GPP component of $X(z)G(z)$ with respect to the polyphase basis $\{F_i'(z)\}$. The proof is complete. $\nabla\,\nabla\nabla$

Eqn. (2.2) gives only the ith GPP component of $x(n) * g(n)$. The convolution output $y(n)$ can be synthesized from the GPP components as follows:

$$Y(z) = \sum_{i=0}^{L-1} F_i'(z)Y_i(z^L) = \sum_{i=0}^{M-1} F_i'(z)\sum_{k=0}^{M-1} X_k(z^{n_k})G_k^{(i)}(z^{n_k}). \tag{2.5}$$

Remarks: Notice that for the uniform case, the corresponding formulas can be obtained by replacing all n_k with M, and p_k with unity. Also notice that even for the nonuniform case, the second-level filter banks with filters $\{H_k'(z)\}$ and $\{F_k'(z)\}$ are constrained to be uniform filter banks with decimation ratio $L = \mathrm{lcm}\{n_k\}$.

Comparison of two-level FB convolver with one-level FB convolver. Theorem 2.1.and 2.2 give us respectively the implementations of one-level and two-level FB convolver as Fig. 2.2(a) and (b). Since $g(n)$ passes through two levels of filter banks, we call the subband convolver in Fig. 2.2(b) two-level FB convolver. From these two figures, it is clear that the two-level FB convolver is a generalization of the one-level FB convolver. By taking $H_k'(z)$ and $F_k'(z)$ to be z^{-i} and z^i respectively, the two-level FB convolver reduce to the one-level FB convolver. As in the one-level FB convolver we see that there is no cross-convolution between subbands in the two-level FB convolver. But in order to obtain the convolution output, the outputs of the subband convolutions have to be interpolated by the synthesis filters $F_k'(z)$ (Fig. 2.2(b)), instead of just interlacing as in the one-level FB biorthonormal convolution (Fig. 2.2(a)). The advantage of the two-level FB convolution theorem over the one-level case will be clear when we discuss the low sensitivity FIR filter structure in Section 4. The two-level FB convolver usually computes the convolution much more accurately than the one-level FB convolver, for the same average bit rate. The complexity of the former is that of the latter plus the cost of an additional synthesis bank $F_k'(z)$ (since $g_k^{(i)}(n)$ can be precomputed and stored). Thus if the complexity of the filter bank $\{F_k'(z)\}$ is low, then the complexity of the new subband convolution is comparable to that of direct convolution.

2.3. Pictorial proof of the subband convolution theorem

The above subband convolution theorems can be proved easily by using a sequence of figures. The pictorial proof of Theorem 2.1 leads us naturally to the two-level FB convolution theorem. It also gives

a clear insight into what is going on in the subbands, and why perfect convolution is perserved when we pass from the one-level FB convolution theorem to the two-level FB convolution theorem. By using the same technique, the subband convolution theorem has been generalized to the most general case of the multidimensional nonuniform filter banks with rational decimation ratios [21].

Consider Fig. 2.3(a), where we want to compute $x(n) * g(n)$. Clearly, any two identity systems $\mathbf{I}_1$ and $\mathbf{I}_2$ can be inserted before and after the filter $G(z)$ without changing the convolution output, as shown in Fig. 2.3(b). If we choose the identity systems to be filter banks with perfect reconstruction, then we can utilize the frequency splitting property of the filter banks and quantize the subband signals according to the energy distribution in each subband. We may also select other identity systems, depending upon the task we want to perform. If we choose $\mathbf{I}_1$ to be the perfect reconstruction system shown in Fig. 1.1(a), and $\mathbf{I}_2$ to be an L-channel delay chain, then we can show that the equivalent system shown in Fig. 2.3(c) is the same as that depicted in Fig. 2.2(a). By using the fact that $L = n_k p_k$, i. e. an L-fold decimator is equivalent to an n_k-fold decimator followed by an p_k-fold decimator, the ith branch of the system in Fig. 2.3(c) (i. e., the system from $x(n)$ to $y_i(n)$) can be redrawn as Fig. 2.3(d). Applying the identity in Fig. 2.3(e), it is clear that the system in Fig. 2.3(d) is equivalent to the ith branch of the system in Fig. 2.2(a). This is the one-level FB biorthonormal convolution theorem that we have described in Theorem 2.1.

Similarly, to prove Theorem 2.2, we select $\mathbf{I}_2$ to be an L-channel biorthonormal filter bank with perfect reconstruction instead of a trivial delay chain. By carrying out exactly the same procedure as above, we can arrive at the result proved in Theorem 2.2.

From the sequence of figures in Fig. 2.3, we see how $x(n)$ and $g(n)$ are split into different subband signals by the two multirate systems. Very similar to the idea of subband coding, the energy distribution of both $x(n)$ and $g(n)$ can be exploited to obtain a more accurate convolution if the computation is done in finite percision.

3. CODING GAIN OF TWO-LEVEL FB CONVOLVERS

In this section, we will consider the coding gain for the quantization of the input signal $x(n)$ only. The filter $g(n)$ is not quantized. For the case of one-level FB orthonormal convolver, the optimal bit allocation and coding gain were discussed in detail in Section 3.2 and 3.3 of [2] respectively for both the uniform and nonuniform cases. For the case of one-level FB biorthonormal convolver, as the formulas for both the optimal bit allocation and coding gain take exactly the same form as (3.16) and (3.17) in [2] respectively. The only difference is that we cannot prove a result similar to Lemma 3.1 and 3.2 in [2], i. e., the coding gain for the biorthonormal convolver cannot be proved to be always greater than unity. So we will not elaborate on the one-level FB convolver. But for the two-level FB convolver, as the subband convolution results are interpolated by the synthesis filters $\{F_k'(z)\}$ as shown in Fig. 2.2(b), the analysis of the output error due to the quantization in the subband is more complicated. Furthermore, the optimal bit allocation scheme is quite different from that of the one-level FB convolver since the energy of $H_i'(z)G(z)$ is different in each

branch in Fig. 2.2(b). We will derive the optimal bit allocation and coding gain formulas for the two-level FB convolver in the rest of this section. Since the uniform convolver is strictly a special case of the nonuniform convolver, we will only derive the result for the nonuniform case. The corresponding formulas for the uniform case follow directly by replacing L and all n_k's with M.

Consider Fig. 2.2(b). Assume that $x(n)$, $g(n)$, and the filter coefficients in all analysis and synthesis filters in the filter bank are real. Then the quantizer operates on real inputs only. Let b_{ki} be the number of bits per sample of $x_k(n)$, allocated to Q_{ki}, the quantizer in the k-th channel in the i-th branch. Therefore the average bit rate is

$$b = \frac{1}{L}\sum_{i=0}^{L-1}\sum_{k=0}^{M-1}\frac{b_{ki}}{n_k}. \tag{3.1}$$

3.1. The noise model

The error due to the quantizer Q_{ki} is defined as

$$q_{ki}(n) \triangleq \widehat{x}_k^{(i)}(n) - x_k(n), \tag{3.2}$$

where $\widehat{x}_k^{(i)}(n)$ is the quantized version of $x_k(n)$ in the i-th branch. The quantization error can be modeled as an additive noise source. Thus the quatizer Q_{ki} can be replaced by the broken line as shown in Fig. 2.2(b). The reason why we choose double index instead of single index as in [2] would be clear after the derivation later.

To analyze the convolution error, we make the following assumptions:

(i) $x(n)$ is zero mean wide sense stationary (WSS) with variance σ_x^2. Then $x_k(n)$ are also WSS, with variance

$$\sigma_{x_k}^2 = \int_0^{2\pi} S_{xx}(e^{j\omega})|H_k(e^{j\omega})|^2\frac{d\omega}{2\pi}, \tag{3.3}$$

where $S_{xx}(e^{j\omega})$ is the power spectrum of $x(n)$.

(ii) $g(n)$ is a deterministic sequence. We define a useful parameter α_{ki}^2 as

$$\alpha_{ki}^2 = M\sum_n |g_k^{(i)}(n)|^2, \tag{3.4}$$

where α_{ki}^2/M can be interpreted as the energy of the subband signal $g_k^{(i)}(n)$.

(iii) $q_{ki}(n)$ is zero mean white with variance $\sigma_{q_{ki}}^2$, where under certain conditions, $\sigma_{q_{ki}}^2$ is related to $\sigma_{x_k}^2$ as

$$\sigma_{q_{ki}}^2 = c\sigma_{x_k}^2 2^{-2b_{ki}}. \tag{3.5}$$

See Chapter 4 of [22] or Appendix C of [3]. Here c is a constant which depends only on the probability distribution of the subband signals $x_k(n)$. We have assumed that c is independent of k which is true only if all $x_k(n)$ have the same probability distribution.

The cross-correlation of $q_{ki}(n)$ is

$$E\{q_{ki}(n)q_{mi}(l)\} = \sigma^2_{q_{ki}}\delta(k-m)\delta(n-l), \tag{3.6a}$$

i. e. $q_{ki}(n)$ is uncorrelated to $q_{mi}(l)$ for $k \neq m$ and for all i, n, l. Notice that $E\{q_{ki}(n)q_{kj}(n)\}$ need not be zero for $i \neq j$. We also assume that $q_{ki}(n)$ is uncorrelated to the subband signals $x_k(l)$, that is

$$E\{x_k(l)q_{ki}(n)\} = 0. \tag{3.6b}$$

3.2. Optimal bit allocation and coding gain for the two-level FB convolver

To derive the optimal bit allocation and coding gain formulas for the two-level FB convolver, we assume that the second set of synthesis filters $F_i'(z)$ are orthonormal (or equivalently the uniform filter banks formed by $\{H_k'(z)\}$ and $\{F_k'(z)\}$ are paraunitary). Consider Fig. 2.2(b). The error in the subband convolution output $y_i(n)$ is

$$q_{y_i}(n) = \sum_{k=0}^{M-1}\Big(q_{ki}(n) * g_k^{(i)}(n)\Big)_{\downarrow p_k}. \tag{3.7}$$

By using (3.4)-(3.6) and the fact that the decimator will not change the variance, the variance of $q_{y_i}(n)$ can be expressed as

$$\begin{aligned}\sigma^2_{q_{y_i}} &= \sum_{k=0}^{M-1}\sigma^2_{q_{ki}}\sum_l |g_k^{(i)}(l)|^2 \\ &= \frac{c}{M}\sum_{k=0}^{M-1} 2^{-2b_{ki}}\sigma^2_{x_k}\alpha^2_{ki}, \quad \text{for } 0 \le i \le L-1.\end{aligned} \tag{3.8}$$

Since the synthesis filters $F_i'(z)$ are orthonormal, the average variance over L samples, $\sigma^2_{q_y}$ of the output error is simply the average of $\sigma^2_{q_{y_i}}$, for $0 \le i \le L-1$, see Section C.4.2 of [3] or [23]. So we have

$$\sigma^2_{q_y} = \frac{1}{L}\sum_{i=0}^{L-1}\sigma^2_{q_{y_i}} = \frac{c}{ML}\sum_{i=0}^{L-1}\sum_{k=0}^{M-1} 2^{-2b_{ki}}\sigma^2_{x_k}\alpha^2_{ki} \tag{3.9}$$

To obtain the optimal bit allocation, we minimize the average output noise variance under the constraint (3.1). We form the Lagrangian

$$\phi = \sigma^2_{q_y} - \lambda\Big(b - \frac{1}{L}\sum_{i=0}^{L-1}\sum_{k=0}^{M-1}\frac{b_{ki}}{n_k}\Big). \tag{3.10}$$

By setting $\partial\phi/\partial b_{ki} = 0$ for all $0 \le i \le L-1$, $0 \le k \le M-1$ and $\partial\phi/\partial\lambda = 0$, we get

$$n_k 2^{-2b_{ki}}\sigma^2_{x_k}\alpha^2_{ki} = D \quad \text{for } 0 \le i \le L-1,\ 0 \le k \le M-1 \tag{3.11}$$

where D is a constant independent of i and k. Let γ_k^2 be the geometric mean of α^2_{ki} over the index i, that is

$$\gamma_k^2 = \prod_{i=0}^{L-1}\left(\alpha^2_{ki}\right)^{1/L}. \tag{3.12}$$

By using (3.1), (3.11) and (3.12), we find that

$$D = 2^{-2b} \prod_{k=0}^{M-1} \left(n_k \sigma_{x_k}^2 \gamma_k^2 \right)^{1/n_k}. \tag{3.13}$$

Substituting (3.13) into (3.11), we find that the optimal number of bits allocated to the quantizer Q_{ki} at the k-th channel in the i-th branch is

$$b_{ki} = b + 0.5 \log_2 \left(n_k \sigma_{x_k}^2 \alpha_{ki}^2 \right) - 0.5 \sum_{j=0}^{M-1} \log_2 \left(n_j \sigma_{x_j}^2 \gamma_j^2 \right)^{1/n_j} \tag{3.14}$$

Periodically time-varying bit allocation. From (3.14), we see that the average bit rate b must be high enough so that $b_{ki} \geq 0$ for all k and i. Notice that b_{ki} are integers and the values evaluated by (3.14) must be rounded off to integers. Intuitively, we would assign more bits to those quantizers in branches where $g(n) * h_i'(n)$ has higher energy and in channels where $x_k(n)$ has higher energy. (3.14) tells exactly how this should be done according to the energy distribution. In the case of the one-level FB convolver, since $g_k^{(i)}(n)$ is simply obtained by time-shifting $g(n)$ (see Fig. 2.2(a)), we would expect that α_{ki}^2 will have very little dependency on i. In this case, b_{ki} are the same for all i and (3.14) reduces to (3.32) in [2]. However in the case of the two-level FB convolver, α_{ki}^2 may differ greatly for different i, especially when the filter $g(n)$ is a frequency selective filter (which is usually the case). Then b_{ki} may vary greatly with respect to i. This is the reason why we use double index. In this case, not all branches are equally important as in the case of the one-level FB convolver, and the coding gain may increase significantly by using this "periodically time-varying" bit allocation scheme.

By using (3.11) and (3.13) and the fact that the filter bank is maximally decimated, i. e. $\sum_{k=0}^{M-1} \frac{1}{n_k} = 1$, we find that the average output noise variance under optimal bit allocation is

$$\sigma_{q_y,opt}^2 = \frac{cD}{M} = \frac{c}{M} 2^{-2b} \prod_{k=0}^{M-1} \left(n_k \sigma_{x_k}^2 \gamma_k^2 \right)^{1/n_k}. \tag{3.15}$$

If $x(n)$ is quantized to b bits, then in the direct convolution the output noise variance due the quantization is found to be

$$\sigma_{\text{direct}}^2 = c 2^{-2b} \sigma_x^2 \sum_l |g(l)|^2. \tag{3.16}$$

Under optimal bit allocation, the coding gain of the two-level FB convolver over the direct form is

$$\begin{aligned} G_{x,\text{ two}} &= \frac{\text{output variance}|_{\text{direct conv}}}{\text{output variance}|_{\text{subband conv}}} \\ &= \frac{\sigma_x^2}{\prod_{i=0}^{M-1} \left(\sigma_{x_i}^2 \right)^{1/n_i}} \times \frac{\sum_n |g(n)|^2}{\frac{1}{M} \prod_{i=0}^{M-1} \left(n_i \gamma_i^2 \right)^{1/n_i}} \end{aligned} \tag{3.17}$$

The "x, *two*" in the subscript in (3.17) indicates that the coding gain is obtained by using the two-level FB convolver and quantizing the signal $x(n)$. This subscript is used to distinguish (3.17) from the deterministic

coding gain which is obtained by quantizing $g(n)$ in the next section. From the right hand side of (3.17), we see that the variation of subband energy of both $x(n)$ and $g(n)$ contributes to the coding gain. The first term is the gain contributed by $x(n)$ and the second term is the gain contributed by $g(n)$.

Summary of crucial assumptions. In the derivation of (3.17), we have assumed that the constant c in (3.16) is the same as that in (3.5) which is true only if $x(n)$ and all $x_k(n)$ have the same probability distribution. Also in the above derivation, we have made used of the orthonormality of the filters $\{F'_k(z)\}$ and the uncorrelated assumptions of $q_{ki}(n)$. Notice that only the *biorthonormality* of the filters $\{F_k(z)\}$ and $\{H_k(z)\}$ is required for (3.17) to be valid, orthonormality of those filters being not necessary. But without the orthonormality of those filters, we *cannot* guarantee that the optimal coding gain in (3.17) is always greater than unity. If the filters $\{F_k(z)\}$ are orthonormal, then we can prove that the coding gain for the two-level FB convolver is always greater than unity, regardless of the *quality* of the filters $\{H_k(z)\}$, $\{F_k(z)\}$, $\{H'_k(z)\}$ and $\{F'_k(z)\}$. Moreover, we can prove that this coding gain is never smaller than that of the one-level FB convolver derived in Section 3.3 in [2], provided that $x(n)$, $g(n)$, $\{H_k(z)\}$ in both cases are the same. More precisely, we have the following lemma:

Lemma 3.1. The coding gain $G_{x,\text{two}}$ of the two-level orthonormal FB convolver (i. e. FB in both levels are orthonormal) is never smaller than that of the one-level orthonormal FB convolver, regardless of the choice of paraunitary filters $\{H'_k(z)\}$, provided that $x(n)$, $g(n)$, $\{H_k(z)\}$ in both cases are the same. Moreover, they are equal if and only if the sequence $g_k^{(i)}(n)$ has the same energy for all $0 \le i \le L-1$. $\diamondsuit$

In [2], it was shown that under optimal bit allocation, the coding gain of the one-level FB convolver is

$$G_{x,\text{one}} = \frac{\sigma_x^2}{\prod_{i=0}^{M-1}\left(\sigma_{x_i}^2\right)^{1/n_i}} \times \frac{\sum_n |g(n)|^2}{\frac{1}{M}\prod_{i=0}^{M-1}\left(n_i\alpha_i^2\right)^{1/n_i}}, \tag{3.18}$$

where α_k^2 is defined as

$$\alpha_k^2 = \frac{M}{L}\sum_{i=0}^{L-1}\sum_n |g_{k,\text{one}}^{(i)}(n)|^2, \tag{3.19}$$

where the "one" in the subscript is used to denote that $g_{k,\text{one}}^{(i)}(n)$ are the subband filters of the one-level FB convolver (see Fig. 3.1). Comparing (3.18) with (3.17), we find that the coding gain formulas for both the one-level and two-level FB convolvers are very similar, except that α_k^2 is replaced by γ_k^2. Therefore in the following proof of Lemma 3.1, we need to establish the relation between α_k^2 and γ_k^2.

Proof of Lemma 3.1. By defining $\mathbf{h}'(z) = [H'_0(z)\ H'_1(z)\ \ldots\ H'_{L-1}(z)]^T$, and $\mathbf{e}(z) = [1\ z^{-1}\ \ldots\ z^{-(L-1)}]^T$, we have $\mathbf{h}'(z) = \mathbf{E}'(z^L)\,\mathbf{e}(z)$, where $\mathbf{E}'(z)$ is the $L \times L$ polyphase matrix of $\mathbf{h}'(z)$. From the definition of $g_k^{(i)}(n)$ and $g_{k,\text{one}}^{(i)}(n)$, it is clear that $g_k^{(i)}(n)$ can be obtained by passing $g_{k,\text{one}}^{(i)}(n)$ through $\mathbf{E}'(z^{p_k})$ as shown in Fig. 3.1. Since $\mathbf{E}'(z)$ is paraunitary, we have [23]

$$\sum_{i=0}^{L-1}\sum_n |g_k^{(i)}(n)|^2 = \sum_{i=0}^{L-1}\sum_n |g_{k,\text{one}}^{(i)}(n)|^2. \tag{3.20}$$

By using (3.4), (3.19) and (3.20), we find the following important equality

$$\alpha_k^2 = \frac{1}{L}\sum_{i=0}^{L-1}\alpha_{ki}^2 = \text{arithmetic mean of } \alpha_{ki}^2. \tag{3.21}$$

By taking the ratio of $G_{x,\text{two}}$ to $G_{x,\text{one}}$, we find that the ratio of the coding gain of the two-level FB convolver to that of the one-level FB convolver is

$$R_x = \prod_{i=0}^{M-1}\left(\frac{\alpha_k^2}{\gamma_k^2}\right)^{1/n_k} \tag{3.22}$$

Using (3.12) and (3.21) and applying the AM-GM inequality, each term in the product in (3.22) is greater or equal to unity with equality if and only if $\alpha_{ki}^2 = \alpha_k^2$ for all i. So we conclude that $R_x \geq 1$, with equality if and only if $\alpha_{ki}^2 = \alpha_k^2 = \gamma_k^2$ for all i. Or equivalently, the sequences $g_k^{(i)}(n)$ have the same energy for all $0 \leq i \leq L-1$. ▽▽▽

Corollary 3.1. $G_{x,\text{two}} \geq 1$ for the two-level orthonormal FB convolver, regardless of the choice of the orthonormal sets of filters $\{H_k(z)\}$ and $\{H_k'(z)\}$. Equality holds if and only if both $\sigma_{x_k}^2$ and $n_k\alpha_{ki}^2$ are independent of k and i. ◇

Proof. This follows directly from the above lemma and Lemma 3.2 in [2]. ▽▽▽

4. LOW SENSITIVITY STRUCTURE FOR FIR FILTERS AND DETERMINISTIC CODING GAIN

Ignore the quantizers in the subbands of $x(n)$ for the discussion of this section. Very similar to the idea of quantizing $x(n)$, we can quantize the filter coefficients $g_k(n)$ in the subbands based on the input signal variance and maximum amplitude of the subband filter coefficients. However, the coefficients have to be treated as deterministic parameters so that overflow is avoided completely. In this implementation, the convolution error due to the coefficient quantization is much smaller than that in the direct form implementation. Let $\widehat{g}_k^{(i)}(n)$ be the quantized version of $g_k^{(i)}(n)$. Then we can redraw Fig. 2.2(a) and (b) as Fig. 4.1(a) and (b) respectively. The implementations in Fig. 4.1(a) and (b) can be regarded as low sensitivity implementations of the filter $g(n)$. In this section, we will discuss in detail first the optimal bit allocation and the coding gain over direct form for the one-level FB convolver and then for the two-level FB convolver. Again we will only derive the formulas for the nonuniform case. The corresponding formulas for the uniform case can be obtained by simply replacing n_k and L with M. For a preview of the advantage of the implementaion, compare Fig. 5.2 and Fig. 5.3. When the same average number of bits is used to quantize the filter coefficients for direct convolution (Fig. 5.2) and subband convolution (Fig. 5.3), the improvement shown in these figures is significant. In the rest of this section, we will translate this improvement into a mathematical formula.

4.1. Low sensitivity FIR filter structures using the one-level FB convolver

With the quantizers inserted in the subbands of $g(n)$ as in Fig. 2.2(a), let b_k be the number of bits per

sample of $g_k(n)$, allocated to the quantizers Q'_k. Then the average bit rate b is defined as:

$$b = \sum_{k=0}^{M-1} \frac{b_k}{n_k} \tag{4.1}$$

4.1.1 The noise model

Define the deterministic quantization error to be

$$q_k^{(i)}(n) \triangleq \widehat{g}_k^{(i)}(n) - g_k^{(i)}(n), \tag{4.2}$$

where $\widehat{g}_k^{(i)}(n)$ is the quantized version of $g_k^{(i)}(n)$. To avoid overflow in the filter coefficients, we assume that the weighting of the most significant bit assigned to the quantizer Q'_k is greater than $g_{k,max}$, where

$$g_{k,max} = \max_{i,n} |g_k^{(i)}(n)|. \tag{4.3}$$

Under this condition, the stepsize in the k-th quantizer would be $\Delta_k = c_1\, g_{k,max} 2^{-b_k}$ and the mean square value of the quantization error $q_k^{(i)}(n)$ is

$$\sigma^2_{q_k^{(i)}} = 1/L_{g_k} \sum_{n=0}^{L_{g_k}-1} |q_k^{(i)}(n)|^2 = c_2\, g^2_{k,max} 2^{-2b_k}, \tag{4.4}$$

where c_1 and c_2 are constants independent of k and i, L_{g_k} is the length of the subband filter $g_k^{(i)}(n)$. In practice, c_1 and c_2 will depend on $g_k^{(i)}(n)$, but the bit allocation and coding gain is insensitive to the variation of these constants. To carry on the analysis, we assume that they are constant. We further assume that:

(i) $x(n)$ is WSS as assumed in the previous section.

(ii) The deterministic cross-correlation of the quantization error $q_k^{(i)}(n)$, approximately satisfies

$$\frac{1}{L_{g_k}} \sum_{p=0}^{L_{g_k}-1} q_k^{(i)}(p)\, q_j^{(i)}(p+m) \approx \sigma^2_{q_k^{(i)}} \delta(k-j)\delta(m). \tag{4.5}$$

This is of course never exact because $q_k^{(i)}(n)$ is FIR.

(iii) The length L_g of $g(n)$ is much greater than that of the analysis filters. So $L_{g_k} \approx L_g/n_k$. This is usually the case if the filter bank is of low complexity.

4.1.2. The optimal bit allocation and the deterministic coding gain

Consider Fig. 4.1(a). The error of the subband convolution output $y_i(n)$, due to quantization of $g_k^{(i)}(n)$, can be expressed as

$$q_{y_i}(n) = \sum_{k=0}^{M-1} \left(x_k(n) * q_k^{(i)}(n) \right)_{\downarrow p_k}. \tag{4.6}$$

By using the assumptions in the noise model and carrying out the exact same procedure in Section 3, we find that the optimal number of bits used to quantize the subband filter $g_k^{(i)}(n)$ is

$$b_k = b + 0.5 \log_2 \sigma^2_{x_k} g^2_{k,max} - 0.5 \sum_{i=0}^{M-1} \log_2 \left(\sigma^2_{x_i} g^2_{i,max} \right)^{1/n_i}. \tag{4.7}$$

Under this optimal bit allocation, the average output variance is

$$\sigma^2_{q_y,opt} = c\,L_g 2^{-2b} \prod_{k=0}^{M-1} \left(\sigma^2_{x_k} g^2_{k,max}\right)^{1/n_k}. \tag{4.8}$$

In contrast, suppose we have convolved directly (i. e., without any filter bank). If $g(n)$ is quantized to b bits and without coefficient overflow, then the output error variance is

$$\sigma^2_{\text{direct}} = c\,L_g 2^{-2b} g^2_{max} \sigma^2_x, \tag{4.9}$$

where $g_{max} = \max_n |g(n)|$.

Therefore, from (4.8) and (4.9), we find that the deterministic coding gain of the one-level FB convolver over the direct form is

$$G_{g,\text{ one}} = \frac{\sigma^2_x}{\prod_{k=0}^{M-1} \left(\sigma^2_{x_k}\right)^{1/n_k}} \times \frac{g^2_{max}}{\prod_{k=0}^{M-1} \left(g^2_{k,max}\right)^{1/n_k}}. \tag{4.10}$$

A lower bound for the coding gain. In the above derivation, orthonormality property of the filter bank is not required, biorthonormality is sufficient. As there is no strong relationship between $g_{k,max}$ and g_{max} , even for the case of orthonormal convolver, the deterministic coding gain cannot be proved to be always greater than unity. The likelihood that the deterministic coding gain is less than unity is very low. In fact, in all the examples we encountered in numerical experiments, the coding gain is quite large. However, if the analysis and synthesis filters have unit energy (this condition indeed implies that the biorthonormal filter bank is orthonormal [24]), we can obtain a (very pessimistic) lower bound for the coding gain. From Appendix A, we have the following very loose relationship between $g_{k,max}$ and g_{max}:

$$g_{k,max} \le \sqrt{L_{H_k}}\, g_{max}, \tag{4.11}$$

where L_{H_k} is the length of the filter $H_k(z)$. Substituting (4.11) into (4.10), we find that the coding gain is lower bounded as

$$G_{g,\text{ one}} \ge \frac{\sigma^2_x}{\prod_{k=0}^{M-1} \left(L_{H_k} \sigma^2_{x_k}\right)^{1/n_k}} \tag{4.12}$$

Remarks: We can also define $g_{ki,max} = \max_n |g^{(i)}_k(n)|$ in the noise model. Based on these parameters, we can minimize the output error variance by using " periodically time-varying" bit allocation scheme (that is, use b_{ki} instead of b_k). Although this is more general than what have been done above, the improvement is negligible for the one-level FB convolver. The reason is that in this convolver, $g^{(i)}_k(n)$ is obtained by time-shifting the input $g(n)$, and $g^{(i)}_{k,max}$ would not vary very much with respect to i. In fact, in all the numerical experiments we carried out, we find that all b_{ki} are the same for all i, even if we allow periodically time-varying bit allocation. However, in the case of the two-level FB convolver, this is not true. The coding gain usually increases by a large amount if a periodically time-varying bit allocation scheme is employed.

4.2. Low sensitivity FIR filter structures using the two-level FB convolver

We can implement FIR filters using the two-level FB convolver instead of the one-level FB convolver. This will give a lower sensitivity (i. e. , provide a much higher deterministic coding gain). Or equivalently, we can afford to quantize the subband filters $g_k^{(i)}(n)$ to a much lower bit rate for a fixed accuracy. The optimal bit allocation and deterministic coding gain will be derived in the following. Again for a preview of the advantage of the two-level FB convolver over the one-level FB convolver, compare Fig. 5.3 and Fig. 5.4. The equivalent filter responses for both the cases are comparable even though the average number of bits used in the two-level FB convolver (2 bits) is only half of that used in one-level FB convolver (4 bits).

Consider Fig. 4.1(b). Let b_{ki} be the number of bits used to quantize the subband filter $g_k^{(i)}(n)$. Then the average bit rate b is defined as in (3.1). The noise model assumed here is the same as that in Section 4.1 except that (4.4) is replaced with

$$\sigma^2_{q_k^{(i)}} = 1/L_{g_k} \sum_{n=0}^{L_{g_k}-1} |q_k^{(i)}(n)|^2 = c_2\, g^2_{ki,max} 2^{-2b_{ki}}, \tag{4.13}$$

where

$$g_{ki,max} \triangleq \max_n |g_k^{(i)}(n)|. \tag{4.14}$$

4.2.1. The optimal bit allocation and the deterministic coding gain

The error at the location $y_i(n)$ in Fig. 4.1(b) can be expressed as (4.6). To carry on the analysis, we will assume that the filter bank $\{F_i'(z)\}$ is paraunitary. By using the same technique as in the previous section, we find that the optimal bit used to quantize $g_k^{(i)}(n)$ is

$$b_{ki} = b + 0.5 \log_2 \sigma^2_{x_k} g^2_{ki,max} - 0.5 \sum_{j=0}^{M-1} \log_2 \left(\sigma^2_{x_j} \beta_j^2\right)^{1/n_j}, \tag{4.15}$$

where

$$\beta_k^2 = \prod_{i=0}^{L-1} \left(g^2_{ki,max}\right)^{1/L} = \text{geometric mean of } g^2_{ki,max}. \tag{4.16}$$

The average output noise variance under optimal bit allocation is

$$\sigma^2_{q_y,opt} = c\, L_g 2^{-2b} \prod_{k=0}^{M-1} \left(\sigma^2_{x_k} \beta_k^2\right). \tag{4.17}$$

From (4.9) and (4.17), we find that the deterministic coding gain of the two-level FB convolver over the direct form is

$$G_{g,\text{ two}} = \frac{\sigma_x^2}{\prod_{k=0}^{M-1} \left(\sigma^2_{x_k}\right)^{1/n_k}} \times \frac{g^2_{max}}{\prod_{k=0}^{M-1} \left(\beta_k^2\right)^{1/n_k}}. \tag{4.18}$$

A lower bound for the coding gain. Again we cannot show that the coding gain is always greater than unity. But by exploiting the result from Appendix A (with $h_k(n)$ replaced with $h_k(n) * h_i'(n)$), we can obtain a (very pessimistic) lower bound similar to (4.12) for the coding gain. The lower bound is

$$G_{g,\text{ two}} \geq \frac{\sigma_x^2}{\prod_{k=0}^{M-1} \left((L_{H_k} + L_{H'} - 1)\sigma^2_{x_k}\right)^{1/n_k}} \tag{4.19}$$

where $L_{H'}$ is the length of the analysis filter $H_i'(z)$, assumed to be the same for all i. By taking the ratio of (4.18) to (4.10), we find that the ratio of the deterministic coding gain of the two-level FB convolver to that of the one-level FB convolver is

$$R_g = \prod_{k=0}^{M-1} \left(\frac{g_{k,max}^2}{\beta_k^2} \right)^{1/n_k}. \tag{4.20}$$

Comparisons of results. Comparing the coding gain formulas in all the cases ($G_{x,\text{one}}$, $G_{x,\text{two}}$, $G_{g,\text{one}}$ and $G_{g,\text{two}}$), we find that all of them have the following form

$$G = \frac{\sigma_x^2}{\prod_{i=0}^{M-1} (\sigma_{x_i}^2)^{1/n_i}} \times \frac{A^2}{\prod_{i=0}^{M-1} (A_i^2)^{1/n_i}}. \tag{4.21}$$

All of them have a common first factor which is always greater than unity when the filter bank is orthonormal. They differ only in the second factor. All of them can be obtained by substituting A^2 and A_i^2 with the corresponding parameters. The only difference is that unlike in the case of the statistical coding gain in Section 3, for the deterministic coding gain we *cannot* prove a result similar to (3.21). That is, we cannot prove that $g_{k,max}^2$ is the arithmetic mean of $g_{ki,max}^2$, even if the filter $H_i'(z)$ is paraunitary. So the ratio of the deterministic coding gain, R_g in (4.20), cannot be proved to be always greater than one. Nevertheless, in practice, we will find that β_k^2 is usually much smaller than $g_{k,max}^2$ for a frequency selective filter $g(n)$. The reason is that under usual situations the arithmetic mean of $g_{ki,max}^2$ would not differ much from $g_{k,max}^2$. But $g_{ki,max}^2$ may vary considerably with respect to i if $g(n)$ is frequency selective. Thus, we may expect that the coding in (4.18) would be much larger than that in (4.10) as we will see in the numerical examples in the following section.

Coding gain when both input signal $x(n)$ and filter $g(n)$ are quantized. When quantizers are inserted in both the subbands of $x(n)$ and $g(n)$, the coding gain is not the product of G_x and G_g. To obtain the coding gain, we apply the optimal bit allocation formulas in (3.14) and (4.15) respectively to the quantization of $x_k(n)$ and $g_k^{(i)}(n)$, and ignore the second order effect. The coding gain is

$$G = \frac{\{\sigma_{q_y,opt}^2\}_x + \{\sigma_{q_y,opt}^2\}_g}{\{\sigma_{\text{direct}}^2\}_x + \{\sigma_{\text{direct}}^2\}_g}, \tag{4.22}$$

where the subscript "x" is used to denote the case when only $x(n)$ is quantized, and "g" is used to denote the case when only $g(n)$ is quantized. We see that the largest error term in (4.22) will dominate the coding gain.

5. NUMERICAL EXAMPLES

In this section, only $g_k(n)$ are quantized, but not $x_k(n)$. In the presence of quantizers in the subbands of $g(n)$, the LTI system with impulse response $g(n)$ is effectively replaced with a periodically time varying system (LPTV) (see Section 10.1 of [3] for an introduction to LPTV system) with period L (see next section for the discussion). To describe the system, we have to characterize all L transfer functions $T_k(z)$ as shown in Fig. 5.1. In all the following examples, we therefore show all transfer functions.

In the first four examples, $g(n)$ is an equiripple lowpass filter with $L_g = 132$. The stopband attenuation $\delta_s = -60$ dB and the passpand ripple size $\delta_p = 0.010$. The frequency responses of $g(n)$ with direct quantization to 4 bits and without quantization are shown in Fig. 5.2, the stopband attenuation reduces to -17 dB and the passband ripple size increases to 0.049 after quantization. In these four examples, we will show the equivalent filters if we implement $g(n)$ by using the low-sensitivity structures (Fig. 4.1(a) and (b)). For comparison of the results in the first four examples, we summary the main features in Table 5.1. In the last two examples, we will show the deterministic coding gain and verify the theoretical values with the experimental values.

Example 1. **4 channel paraunitary (PU) filter bank (one-level FB convolver):** $L = M = 4$, and $b = 4$ bits. The 4 channel filter bank in Fig. 4.1(a) is taken to be a tree-structured PU filter bank obtained by using two-channel PU filter bank in a symmetric tree. The two-channel PU system uses Filter 8A in [25]. If we implement the analysis bank $\{H_k(z)\}$ in lattice form, we need only 8 multiplications per input sample. The corresponding optimal bit allocation is $b_0 = 10$, $b_1 = 5$, $b_2 = 1$, $b_3 = 0$ bits. As shown in Fig. 5.3, the stopband attenuations of all the 4 filters $T_i(z)$ are more than 42 dB, i. e. more than 25 dB better than that of the direct quantization. The passband ripple $\delta_p = 0.013$. The effect of quantization on the ripple size is negligible.

Example 2. **4 channel PU filter bank (two-level FB convolver):** $L = M = 4$, and $b = 2$ bits. Both the filter banks formed by $\{H_k(z)\}$ and $\{H'_i(z)\}$ are taken to be the filter bank used in Example 1. The corresponding bit allocation is shown in Table 5.2. As we would expect, b_{ki} are large for $i = 0$ because most of the energy of $G(z)$ is in the first branch. As shown in Fig. 5.4, the stopband attenuations (44 dB) are comparable to that obtained in Example 1 but the average bit rate b is reduced to half. The passband ripple $\delta_p = 0.015$.

Example 3. **4 and 8 channel DCT coders (one-level FB convolver):** $b = 4$ bits, we use the DCT filter bank, shown in Fig. 4.1 in [2]. In a transform coder filter bank, the polyphase matrix $\mathbf{E}(z)$ of the analysis filters is a constant matrix $\mathbf{T}$. In this example, two cases of $\mathbf{T}$ are considered: (i) 4×4 DCT matrix (ii) 8×8 DCT matrix as defined in Eq. (12.157) [22]. DCT has the advantage that the analysis filters have linear phase and there exists fast algorithm for the computation of DCT. The corresponding bit allocations are shown in Table 5.3. For each case, we show only one transfer function $T_0(z)$ in Fig. 5.5 for simplicity. We see that for $M = 4$, the stopband attenuation is 32 dB and $\delta_p = 0.022$. For $M = 8$, the stopband attenuation is 38 dB and $\delta_p = 0.012$.

Example 4. **4 and 8 channel DCT coders (two-level FB convolver):** $b = 2$ bits. The filter bank used here is the same as Example 3. And $\{H'_k(z)\}$ is identical to $\{H_k(z)\}$. The optimal bit allocation for 4×4 DCT is shown in Table 5.4. The corresponding optimal bit allocation for 8×8 DCT is shown in Table 5.5. For simplicity, we show only $T_0(z)$ in Fig. 5.6. The stopband attenuations for $M = 4$ and $M = 8$ are 27 dB and 33 dB respectively. The passband ripple size increases to 0.035 and 0.017 respectively for $M = 4$

and $M = 8$.

Example 5. **Coding gain (one-level FB convolver):** $M = 4$ and $b = 8$ bits. The filter bank used here is the same as that used in Example 1. The input signal $x(n)$ is taken to be an AR(5) process with autocorrelation coefficients $R(k)$ obtained from Table 2.2 of [22] (lowpass speech source). The first two rows of Table 5.6 show respectively the coding gain obtained from (4.10) ($G_{g,\text{one}}$) and that obtained from experiment ($G_{g,\ expt,\text{one}}$) for 5 different filters $g(n)$ (Filter 1 is the $g(n)$ used in the previous 4 examples). In most cases the theoretical value obtained from (4.10) is very close to the experimental result, in spite of the many statistical assumptions used.

Example 6. **Coding gain (two-level FB convolver):** The filter bank formed by $\{H_i'(z)\}$ is identical to that formed by $\{H_k(z)\}$ and other conditions are the same as those in Example 5. The coding gain obtained from (4.18) ($G_{g,\text{two}}$) and that obtained from experiment ($G_{g\ expt,\text{two}}$) for the same set of 5 different filters $g(n)$ are shown in the third and fourth rows of Table 5.6 respectively. Again we see that the theoretical values are very close to the experiment results. The performance of the two-level FB convolvers is much better (8.7–17.4 dB or equivalently 1.5–3 bits approximately) than that of one-level FB convolvers for all the 5 cases. The ratios of the coding gain for the two-level FB convolver to that of the one-level FB convolver, R_g (theoretical) and $R_{g,expt}$ (experimental) are shown in the last two rows of Table 5.6.

From the first four examples, we notice that the performance of the DCT coder is not as good as that of the PU filter bank transformer in Example 1 and 2. The reason is that the analysis filters of the DCT coder have a smaller stopband attenuation. The leakage from the adjacent band is quite large. In the last two examples, we see that the coding gain for the two-level FB convolver is much larger than that of the one-level FB convolver although we cannot prove theoretically that this is always true. By using the two-level FB convolvers in the convolution, we get a much higher accuracy at the expense of the cost of one filter bank.

6. RELATION TO BLOCK FILTER AND ALIASING EFFECT

6.1. Convolvers in the view of block filter

It is well-known [11], [12], [13], [6], [Chapter 10,[3]] that block filtering is a technique to implement a scalar filter $G(z)$ in such a way as to increase the parallelism. In this section, we will explore the relationship between the filter bank convolver and the conventional block filtering technique. It was shown in [26] that the nonuniform system of Fig. 1.1 can be expanded as an L-channel uniform system. The M pairs of filters $\{H_k(z),\ F_k(z)\}$ in the nonuniform system are replaced by the L pairs of filters, say $\{H_k''(z),\ F_k''(z)\}$, in the uniform system. We will discuss the uniform case only, as the nonuniform problem can be translated to uniform case.

6.1.1. Conventional block filtering

Given any scalar filter $G(z)$, we can implement it by using block filtering technique as shown in

Fig. 6.1(a). The matrix $\mathbf{G}(z)$ in Fig. 6.1(a) is a pseudocirculant matrix and it can be written as:

$$\mathbf{G}(z) = \begin{pmatrix} G_0(z) & G_1(z) & G_2(z) & \dots & G_{M-1}(z) \\ z^{-1}G_{M-1}(z) & G_0(z) & G_1(z) & \dots & G_{M-2}(z) \\ z^{-1}G_{M-2}(z) & z^{-1}G_{M-1}(z) & G_0(z) & \dots & G_{M-3}(z) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z^{-1}G_1(z) & z^{-1}G_2(z) & z^{-1}G_3(z) & \dots & G_0(z) \end{pmatrix}, \tag{6.1}$$

where $G_i(z)$ is the ith polyphase component of the scalar filter $G(z)$. In fact, the multirate system in Fig. 6.1(a) is a linear time-invariant (LTI) system if and only if $\mathbf{G}(z)$ is a pseudocirculant matrix [27]. From (6.1), it is clear that we have the following relationship between $[\mathbf{G}(z)]_{ik}$, the elements of the matrix $\mathbf{G}(z)$ and the filter $G(z)$:

$$z^{-i}G(z) = \sum_{k=0}^{M-1} z^{-k}\left[\mathbf{G}(z^M)\right]_{ik}. \tag{6.2}$$

Moreover, it can be shown [3] that the matrix $\mathbf{G}(z)$ is paraunitary if and only if the filter $G(z)$ is an allpass filter. When $G(z)$ is FIR, this is impossible unless $G(z)$ is a delay.

6.1.2. Relation of one-level FB convolver to conventional block filtering

In the case of one-level FB convolver with uniform decimation ratios, the ith Type 2 polyphase component of $x(n) * g(n)$ can be written as (2.1) with $L = M$ and all $p_k = 1$. We reproduce the equation here for convenience:

$$Y_i(z) = \left(z^{-i}X(z)G(z)\right)_{\downarrow M} = \sum_{k=0}^{M-1} X_k(z)G_k^{(i)}(z), \quad 0 \le i \le M-1, \tag{6.3}$$

where the subband signal $G_k^{(i)}(z)$ is the kth GPP component of $z^{-i}G(z)$ with respect to $\{H_l(z)\}$ as defined in Fig. 1.1(b). By writing (6.3) for all values of i, we obtain the matrix equation:

$$\begin{pmatrix} Y_0(z) \\ Y_1(z) \\ \vdots \\ Y_{M-1}(z) \end{pmatrix} = \mathbf{G}_{\text{one}}(z)\mathbf{x}(z), \tag{6.4}$$

where the column vector $\mathbf{x}(z) = \begin{bmatrix} X_0(z) & X_1(z) & \dots & X_{M-1}(z) \end{bmatrix}^T$ and the matrix $\mathbf{G}_{\text{one}}(z)$ is defined as:

$$\mathbf{G}_{\text{one}}(z) = \begin{pmatrix} G_0^{(0)}(z) & G_1^{(0)}(z) & \dots & G_{M-1}^{(0)}(z) \\ G_0^{(1)}(z) & G_1^{(1)}(z) & \dots & G_{M-1}^{(1)}(z) \\ \vdots & \vdots & \ddots & \vdots \\ G_0^{(M-1)}(z) & G_1^{(M-1)}(z) & \dots & G_{M-1}^{(M-1)}(z) \end{pmatrix}. \tag{6.5}$$

By the definition of Type 2 polyphase representation, the output of the convolution $y(n)$ can be written compactly as:

$$Y(z) = \tilde{\mathbf{e}}(z) \begin{pmatrix} Y_0(z^M) \\ Y_1(z^M) \\ \vdots \\ Y_{M-1}(z^M) \end{pmatrix} = \tilde{\mathbf{e}}(z)\mathbf{G}_{\text{one}}(z^M)\mathbf{x}(z^M), \tag{6.6}$$

where $\tilde{\mathbf{e}}(z)$ is the row vector $\begin{bmatrix}1 & z & \dots & z^{M-1}\end{bmatrix}$. From (6.6), we immediately get the implementation in Fig. 6.1(b), by using the fact that $X_k(z) = \left[X(z)H_k(z)\right]_{\downarrow M}$.

Comparison between one-level FB convolver and conventional block filtering. Comparing Fig. 6.1(a) and (b), we discover that the one-level FB convolver is a generalized version of block filtering. Instead of decomposing $x(n)$ and $g(n)$ into their conventional polyphase components as we did in Section 6.1.1, we decompose $x(n)$ and $g(n)$ into their GPP components respectively with respect to two separate sets of polyphase basis, namely $\{H_i(z)\}$ and $\{F_i(z)\}$. The delay chain before the block filter is replaced by a more general analysis bank with filters $\{H_k(z)\}$. Therefore, we can view the convolver as a generalized block filtering technique, which provides not only the advantage of parallelism, but also the advantage of coding gain when implemented in finite percision. Of course, the coding gain is obtained at the expense of the cost of one filter bank. This generalized block filtering technique provides a good tradeoff between the coding gain and the complexity. This is the advantage that the conventional block filtering technique does not have. By using GPP representation, a relationship similar to (6.2) between $\left[\mathbf{G}_{\text{one}}(z)\right]_{ik}$ and $G(z)$ can be interpreted nicely as:

$$z^{-i}G(z) = \sum_{k=0}^{M-1} \left[\mathbf{G}_{\text{one}}(z^M)\right]_{ik} H_k(z). \tag{6.7}$$

It also can be proved (see Appendix B) that the matrix $\mathbf{G}_{\text{one}}(z)$ is paraunitary if and only if the filter $G(z)$ is an allpass function, provided that the set of filters $\{H_k(z)\}$ is paraunitary.

6.1.3. Relation of two-level FB convolver to conventional block filtering

For two-level FB convolver with uniform decimation ratios, the ith GPP component of $x(n) * g(n)$ with respect to the polyphase basis $\{F_i'(z)\}$ is:

$$Y_i(z) = \sum_{k=0}^{M-1} X_k(z) G_k^{(i)}(z), \quad 0 \le i \le M-1, \tag{6.8}$$

where the subband signals $G_k^{(i)}(z)$ are the kth GPP components of $H_i'(z)G(z)$ with respect to $\{H_l(z)\}$ as defined in Fig. 2.1. By writing (6.8) for all values of k, we get the equations similar to (6.4) and (6.5), except that the matrix $\mathbf{G}_{\text{one}}(z)$ is replaced by $\mathbf{G}_{\text{two}}(z)$, where $\left[\mathbf{G}_{\text{two}}(z)\right]_{ki} = G_k^{(i)}(z)$. By defining the row vector $\mathbf{f}'(z) = \begin{bmatrix}F_0'(z) & F_1'(z) & \dots & F_{M-1}'(z)\end{bmatrix}$, the output of the convolution $y(n)$ can be reconstructed from the GPP components $Y_i(z)$ (as defined in (6.8)) as:

$$Y(z) = \sum_{k=0}^{M-1} Y_i(z^M) F_k'(z) = \mathbf{f}'(z)\mathbf{G}_{\text{two}}(z^M)\mathbf{x}(z^M), \tag{6.9}$$

where the column vector $\mathbf{x}(z)$ is as defined in previous section. From (6.9), we get the implementation of the two-level FB convolver as in Fig. 6.1(c).

Comparison between one- and two-level FB convolver in the light of block filtering. Comparing Fig. 6.1(b) and (c), clearly the two-level FB convolver is a generalized version of one-level FB convolver.

In the two-level FB convolver, the "advance chain" in the one-level FB convolver after the block filter is replaced by a more general synthesis bank with filters $\{F_i'(z)\}$. Notice that the sets of filters $\{H_k(z)\}$ and $\{F_i'(z)\}$ can come from two different biorthonormal systems. The relationship between $[\mathbf{G}_{\text{two}}(z)]_{ik}$ and $G(z)$ can be written as:

$$H_i'(z)G(z) = \sum_{k=0}^{M-1} [\mathbf{G}_{\text{two}}(z^M)]_{ik} H_k(z). \tag{6.10}$$

Similarly, we can prove (Appendix B) that the matrix $\mathbf{G}_{\text{two}}(z)$ is paraunitary if and only if the filter $G(z)$ is an allpass function, provided that the sets of filters $\{H_k(z)\}$ and $\{H_k'(z)\}$ are paraunitary.

Remarks: The adaptive structure in Fig. 2 of [14] is a simplified version of the two-level FB convolver in Fig. 6.1(c) (with the second set of filter $F_k'(z)$ replaced with $F_k(z)$).

6.2. Aliasing effects and the equivalent LPTV filter in the presence of quantizers

In the presence of quantizers in the subband of $g(n)$, the equivalent system is not a linear time invariant (LTI) system anymore. We will discuss the aliasing effect and the relation between the subband convolver and a linear periodically time varying (LPTV) system.

Let $Q_k^{(i)}(z)$ be the z transform of $q_k^{(i)}(n)$, where $q_k^{(i)}(n)$ is defined in (4.2). Define the matrix $\mathbf{Q}(z)$

$$\mathbf{Q}(z) = \begin{pmatrix} Q_0^{(0)}(z) & Q_1^{(0)}(z) & \dots & Q_{M-1}^{(0)}(z) \\ Q_0^{(1)}(z) & Q_1^{(1)}(z) & \dots & Q_{M-1}^{(1)}(z) \\ \vdots & \vdots & \ddots & \vdots \\ Q_0^{(M-1)}(z) & Q_1^{(M-1)}(z) & \dots & Q_{M-1}^{(M-1)}(z) \end{pmatrix}. \tag{6.11}$$

Let $\widehat{\mathbf{G}}_{\text{one}}(z)$ be the quantized version of $\mathbf{G}_{\text{one}}(z)$. Then $\widehat{\mathbf{G}}_{\text{one}}(z) = \mathbf{G}_{\text{one}}(z) + \mathbf{Q}(z)$. The system in Fig. 6.1(b) can be drawn equivalently as that in Fig. 6.2(a). The upper path gives the desired output and the lower path represents the error. By using the polyphase representation, Fig. 6.2(a) can be redrawn as Fig. 6.1(b) where

$$\mathbf{P}(z) = \mathbf{Q}(z)\,\mathbf{E}(z), \tag{6.12}$$

and $\mathbf{E}(z)$ is the polyphase matrix of the analysis filters $H_k(z)$. From Fig 6.2, we see that the lower path is an LPTV filter and it is an LTI filter if and only if the matrix $\mathbf{P}$ is psuedocirculant ([3], Section 10.1). For the case of the two-level FB convolver as in Fig. 6.1(c), the similar result holds except that the matrix $\mathbf{P}(z)$ is replaced by

$$\mathbf{P}(z) = \mathbf{R}'(z)\,\mathbf{Q}(z)\,\mathbf{E}(z), \tag{6.13}$$

where $\mathbf{R}'(z)$ is the Type 2 polyphase matrix of the synthesis filters $F_i'(z)$.

Let $\mathbf{d}(z) = [\, d_0(z) \;\; d_1(z) \;\; \dots \;\; d_{M-1}(z)\,]^T = \mathbf{P}(z^M)\,\mathbf{e}(z)$, where $\mathbf{e}(z) = [1 \;\; z^{-1} \;\; \dots \;\; z^{-(M-1)}]^T$ and let $T_i(z) = z^{-i}G(z) + d_i(z)$. Then the system in Fig. 6.2 can be redrawn as Fig. 5.1. The aliasing components $A_i(z)$ (see Eq. (5.4.7) of [3]) can be expressed as:

$$A_i(z) = \frac{1}{M} \sum_{k=0}^{M-1} z^k d_k(zW^i), \quad \text{for} \;\; 1 \le i \le M-1 \tag{6.14}$$

and

$$A_0(z) = G(z) + \frac{1}{M}\sum_{k=0}^{M-1} d_k(z). \tag{6.15}$$

$G(z)$ is the desired response, $\frac{1}{M}\sum_{k=0}^{M-1} d_k(z)$ represents the distortion and for $1 \le i \le M-1$, $A_i(z)$ are the aliasing components. The error due to the aliasing and distortion can be written as

$$\begin{aligned}
\varepsilon^2 &= \sum_{i=1}^{M-1} \left|A_i(e^{j\omega})\right|^2 + \left|A_0(e^{j\omega}) - G(e^{j\omega})\right|^2 \\
&= \sum_{i=0}^{M-1} \left|\frac{1}{M}\sum_{k=0}^{M-1} e^{jk\omega} d_k(e^{j(\omega-2\pi i/M)})\right|^2 \\
&\le \frac{1}{M^2}\sum_{i=0}^{M-1}\sum_{k=0}^{M-1} |d_k(e^{j(\omega-2\pi i/M)})|^2
\end{aligned} \tag{6.16}$$

The magnitude responses of $d_k(z)$ for Example 1 are shown in Fig. 6.3. All the magnitude responses are under 40 dB even though the coefficients are quantized to an average bit rate of 4 bits only.

6.3. Subband implementation of LPTV filters

From the earlier discussion in this section, it is natural to ask if the subband convolver can be modified to implement an LPTV filter. In the following we will show that the answer is in the affirmative. The implemenation leads to a low sensitivity structure for LPTV filters.

Given an LPTV filter with period L, we can characterize the filter by a set of L transfer functions $\{T_i(z)\}$ as shown in Fig. 5.1. Notice from the figure that the ith polyphase component $y_i(n)$ of the output of the LPTV filter is completely determined by the transfer function $T_i(z)$. By Theorem 2.1, $y_i(n)$ can be obtained as:

$$y_i(n) = \Big(x(n) * t_i(n)\Big)_{\downarrow L} = \sum_{k=0}^{M-1} \Big(x_k(n) * t_{ik}(n)\Big)_{\downarrow p_k}, \tag{6.17}$$

where $t_i(n)$ is the impulse response of $T_i(z)$, $x_k(n)$ are defined in Fig. 1.1(a), and $t_{ik}(n)$ are the subband signals obtained by replacing $g(n-i)$ in Fig. 1.1(b) with $t_i(n)$. The periodically time varying bit allocation can be employed to achieve a low sensitivity structure for LPTV filters.

7. LOW SENSITIVITY STRUCTURES FOR IIR FILTERS

If we consider the IIR filtering problem as two FIR filtering problems (one in the forward path and one in the feedback path), then the low sensitivity structure for FIR filters can be applied to obtained a low sensitivity structure for IIR filters. Though the application is straight forward, several issues like stability and causality must be taken care. Any filter with rational transfer function can be considered as a cascade of an all-zero filter and an all-pole filter. In Section 4, we have already considered the structure for all-zero filters. Thus we will study the low sensitivity structure only for an all-pole filter. Let

$$G(z) = \frac{1}{1 + b(1)z^{-1} + \ldots + b(N)z^{-N}}. \tag{7.1}$$

Then the output of the convolution $x(n) * g(n)$ can be written as

$$y(n) = -\sum_{k=1}^{N} b(k)y(n-k) + x(n). \tag{7.2}$$

The summation at the right hand sight can be considered as the convolution of $b(n)$ and $y(n)$. This gives the implementation in Fig. 7.1. If the spectrum of $x(n)$ is known, then so is that of $y(n)$. The coefficients $b(n)$ can be quantized in the subbands according to the energy distribution of $y(n)$ and $b(n)$. Thus the output error due to quantization can be minimized by optimal bit allocation. However, the low sensitivity structure for IIR filters is not as useful as that for FIR filters because of the following reasons:

Stability. From Section 6, we know that the equivalent system with subband quantization is an LPTV system. With an LPTV system in a feedback loop as in Fig. 7.1, the analysis of stability of the overall filter is very difficult [28]. Even though the original filter $g(n)$ is stable, we cannot ensure the stability of the resultant filter after quantization.

Causality. Assume that $H_k(z)$ and $F_k(z)$ are causal filters. If the system in Fig. 1.1 is perfect reconstruction system, then the M-th channel filter bank will introduce a delay of at least $M-1$ samples [8]. That means that the subband convolver will introduce some delay, say D. At the instant of the computation of $y(n)$, the output of the convolver in the feedback path is $y(n-D) * b(n-D)$. At the time instant n, to compute the summation in the right hand side of (7.2), we need $y(n+i)$ for $0 < i \leq D$, which is impossible. To avoid the noncausality, we can use (7.2) repeatedly to get

$$y(n) = -\sum_{k=D+1}^{N+D} c(k)y(n-k) + \sum_{k=0}^{D} d(k)x(n-k). \tag{7.3}$$

However, D more multipliers ($d(k)$ in the second summation) are required in the implementation. If the delay D is long, then this computational overhead is large.

Complexity. As we mentioned earlier, for the subband convolver to be useful, the FIR filter must have a long impulse response. But for the IIR filter, the filter seldom has an order N greater than 10. In this case, the complexity of the filter bank is comparable to (or even higher than) that of the filter $g(n)$.

Frequency selectivity. Let $B(z) = b(1)z^{-1} + b(2)z^{-2} + \ldots + b(N)z^{-N}$. The deterministic coding of the subband convolver is high when the FIR filter $B(z)$ is frequency selective. But since $B(z)$ is only a part of the denominator of an IIR filter, the energy of $B(z)$ is distributed all over the frequency domain, even when the filter $g(n)$ is a frequency selective filter.

8. CONCLUDING REMARKS

8.1. IFIR filter as a special case of subband convolver

IFIR filters were introduced in [15] to design narrowband filters. In lowpass case, if the stopband edge is smaller than π/M, then $G(z)$ can be approximated by a cascade of two filters as:

$$G(z) \approx G^{(0)}(z^M)I(z), \tag{8.1}$$

where $I(z)$ is a low cost filter. The number of coefficients in $G^{(0)}(z)$ is roughly equal to $1/M$ of that in $G(z)$. Fig. 8.1(a) shows the implementation of an IFIR filter.

From Fig. 1.1(b), $G(z)$ can be decomposed into its GPP components as

$$G(z) = \sum_{k=0}^{M-1} G_k^{(0)}(z^M) H_k(z). \tag{8.2}$$

The decomposition is exact. Fig. 8.1(b) shows the implementaion. If $G(z)$ has passband smaller than π/M, then only $G_0^{(0)}(z)$ in (8.2) has significant energy. By dropping all the other unimportant channels in Fig. 8.1(b) corresponding to $G_k^{(0)}(z)$ ($k = 1, 2, \ldots, M-1$), Fig. 8.1(b) reduces to Fig. 8.1(a) (with $H_0(z)$ and $F_0(z)$ regarded as $I(z)$ and $J(z)$ respectively). Therefore, more generally, if $G(z)$ is a multiband filter, the subband convolver can be used to approximate $G(z)$ by retaining the channels which contain most of the energy.

8.2. Conclusions and open problems

In this paper, we have generalized the subband convolution theorem in [2]. We have derived the coding gain for the generalized convolver, and it was proved that this coding gain is always greater than that of the one-level FB convolver in [2]. We also unified the subband convolvers, GPP representation, block filtering, LPTV filters, and IFIR filters under one framework. This framework provides us a better understanding of the subband convolvers. As an application of the convolution theorem, a low sensitivity structure for FIR filters is proposed. We have defined the deterministic coding gain of the low sensitivity structure and demonstrated that the coding gain is high. Even when the filter coefficients are quantized to a very low bit rate, we can get filters of small passband ripple and large stopband attenuation. However, the linear phase property of the filter $g(n)$ is generally not preserved when we implement the filter in the low sensitivity structure. How to exploit the symmetry (or some other properties like Mth-band property and etc) of the filter to save computation is still an open problem. Another important issue which is not being addressed in this paper is how to design an optimal filter bank to maximize the convolutional coding gain. Even for the case of traditional transform coder, the solution is unknown.

Appendix A. Proof of (4.11)

First, we will prove a fact about vector norms inequality. This will be used to derive (4.11).

Fact B.1. Let $\mathbf{v} = [v_0\ v_1\ \ldots\ v_{N-1}]^T$, and let $\|\cdot\|_1$ and $\|\cdot\|_2$ denote 1-norm and 2-norm respectively. Then

$$\|\mathbf{v}\|_1 \leq \sqrt{N}\,\|\mathbf{v}\|_2 \tag{A.1}$$

Proof. Let $\mathbf{1} = [1\ 1\ \ldots\ 1]^T$ and $\mathbf{u} = [|v_0|\ |v_1|\ \ldots\ |v_{N-1}|]^T$. Then we have

$$\begin{aligned}\|\mathbf{v}\|_1 &= \sum_{i=0}^{N-1} |v_i| = \mathbf{1}^T \cdot \mathbf{u} \\ &\leq \|\mathbf{1}\|_2 \|\mathbf{u}\|_2 = \|\mathbf{1}\|_2 \|\mathbf{v}\|_2 = \sqrt{N}\,\|\mathbf{v}\|_2,\end{aligned} \tag{A.2}$$

using Cauchy-Schwarz inequality. ▽▽▽

By definition, we have

$$\begin{aligned}g_{k,max} &= \max_{i,n} |g_k^{(i)}(n)| = \max_{i,n} \left|\left(g(n-i) * h_k(n)\right)_{\downarrow n_k}\right| \\ &= \max_{n} |g(n) * h_k(n)| \\ &\leq g_{max} \sum_{n=0}^{L_{H_k}-1} |h_k(n)|\end{aligned} \tag{A.3}$$

The third equality follows from the fact that $\left(g(n-i) * h_k(n)\right)_{\downarrow n_k}$ is one of the polyphase components of $g(n) * h_k(n)$. The last inequality follows directly from triangular inequality. Applying Fact B.1 and the fact that $h_k(n)$ has unit energy (2-norm is unity), (4.11) follows immediately.

Appendix B. Proof of some facts in block filtering

Lemma B.1. One-level FB convolver. Suppose that the set of filters $\{H_k(z)\}$ is paraunitary. Then the matrix $\mathbf{G}_{\text{one}}(z)$ defined in Fig. 6.1(b) is paraunitary if and only if the filter $G(z)$ is an allpass function. ◇

Proof. By writing (6.7) for all values of $i = 0,\ 1,\ \ldots,\ M-1$, we have the following matrix equation:

$$\begin{pmatrix} G(z) \\ z^{-1}G(z) \\ \vdots \\ z^{-M+1}G(z) \end{pmatrix} = \mathbf{G}_{\text{one}}(z^M) \begin{pmatrix} H_0(z) \\ H_1(z) \\ \vdots \\ H_{M-1}(z) \end{pmatrix} = \mathbf{G}_{\text{one}}(z^M)\mathbf{E}(z^M)\mathbf{e}(z), \tag{B.1}$$

where the matrix $\mathbf{E}(z)$ is the polyphase matrix of the filters $\{H_k(z)\}$ and $\mathbf{e}(z) = \begin{bmatrix}1 & z^{-1} & \ldots & z^{-M+1}\end{bmatrix}^T$. Substituting z with zW^{-i} for $i = 0,\ 1,\ \ldots,\ M-1$ into the above equation, we get

$$\mathbf{\Lambda}(z)\mathbf{W}\mathbf{\Phi}_G(z) = \mathbf{G}_{\text{one}}(z^M)\mathbf{E}(z^M)\mathbf{\Lambda}(z)\mathbf{W}, \tag{B.2}$$

where $\mathbf{\Lambda}(z)$ is the diagonal matrix $diag\begin{bmatrix}1 & z^{-1} & \ldots & z^{-M+1}\end{bmatrix}$, $\mathbf{\Phi}_G(z) = diag\begin{bmatrix}G(z) & G(zW) & \ldots & G(zW^{M-1})\end{bmatrix}$ and $\mathbf{W}$ is the $M \times M$ DFT matrix with $\left[\mathbf{W}\right]_{ki} = W^{ki}$. Since $\mathbf{\Lambda}(z)$, $\mathbf{W}$ and $\mathbf{E}(z)$ are paraunitary matrices, $\mathbf{G}_{\text{one}}(z)$ is paraunitary if and only if $\mathbf{\Phi}_G(z)$ is. The proof is complete. ▽▽▽

Lemma B.2. Two-level FB convolver. Suppose that the sets of filters $\{H_k(z)\}$ and $\{H'_k(z)\}$ are paraunitary. Then the matrix $\mathbf{G}_{\text{two}}(z)$ defined in Section 6.1.3 (Fig. 6.1(c)) is paraunitary if and only if the filter $G(z)$ is an allpass function. $\Diamond$

Proof. The proof is similar to that of Lemma B.1. By using (6.10) and following the procedure in the proof above, the lemma follows. $\triangledown\triangledown\triangledown$

References

[1] R. E. Blahut, *Fast algorithms for digital signal processing*, Addison-Wisley, 1985.

[2] P. P. Vaidyanathan, "Orthonormal and biorthonormal filter-banks as convolvers, and convolutional coding gain," *IEEE Trans. on Signal Processing*, June 1993.

[3] P. P. Vaidyanathan, *Multirate systems and filter banks*, Englewood Cliffs, NJ: Prentice Hall, 1993.

[4] M. Vetterli, and C. Herley,"Wavelets and filter banks," *IEEE Trans. on Signal Processing*, Vol. 40, no. 9, pp. 2209-2232, Sep. 1992.

[5] M. J. T. Smith, and T. P. Barnwell, "A new filter-bank theory for time-frequency representation," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, pp. 314-327, March 1987.

[6] P. P. Vaidyanathan,"Multirate digital filters, filter banks, polyphase networks, and applications: a tutorial," *Proc. of the IEEE*, vol 78, pp. 56-93, Jan. 1990.

[7] A. N. Akansu and Y. liu, "On signal decomposition techniques," *Optical engr.*, pp. 912-920, July 1991.

[8] M. Vertteli, "A theory of multirate filter banks," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, pp. 356-372, Mar. 1987.

[9] S. K. Chan, and L. R. Rabiner,"Analysis of quantization errors in the direct form for finite impulse response digital filters," *IEEE Trans. on Audio and Electroacoustics*, pp. 3354-366, Aug. 1973.

[10] A. Steffen, *Digital pulse compression using multirate filter banks*, Hartung-Gorre Verlag, 1991.

[11] C. S. Burrus, "Block implementation of digital filters," *IEEE Trans. on Circuit Theory*, vol. CT-18, pp. 697-701, Nov. 1971.

[12] S. K. Mitra, and R. Gnanasekaran, "Block implementation of recursive digital filters: new structures and properties," *IEEE Trans. on Circuits and Systems*, vol. CAS-25, pp. 200-207, April 1978.

[13] C. W.Barnes, and S. Shinnaka, "Block-shift invariance and block implementaion of discrete-time filters," *IEEE Trans. on Circuits and Systems*, vol. CAS-27,pp. 667-672, Aug. 1980.

[14] A. Gilloire, and M. Vetterli, " Adaptive filtering in subbands with critical sampling: Analysis, experiments and application to acoustic cancellation," *IEEE Trans. on Signal Processing*, vol. 40, pp. 1862-75, Aug. 1992.

[15] Y. Neuvo, C. -Y. Dong, and S. K. Mitra, "Interpolated finite impulse response filters," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. ASSP-32, pp. 563-570, June 1984.

[16] M. Vetterli,"Running FIR and IIR filtering using multirate filter banks," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, pp. 730-738, Jan. 1988.

[17] R. E. Crochiere, and L. R. Labiner, *Multirate Digital Signal processing*, Englewood Cliffs, NJ: Prentice Hall, 1983.

[18] M. Bellanger, G. Bonnerot, and M. Coudreuse, "Digital filtering by polyphase network: application to sample rate alteration and filter banks," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. 24,

pp. 109-114, April 1976.

[19] A. K. Soman, and P. P. Vaidyanathan, "Generalized polyphase representation in multirate signal processing," *IEEE Trans. on Circuits and Systems*, submitted.

[20] A. K. Soman, and P. P. Vaidyanathan, "On orthonormal wavelets and paraunitary filter banks," *IEEE Trans. on Signal Processing*, vol. 41, March 1993.

[21] T. Chen, and P. P. Vaidyanathan, "Vector space framework for unification of one- and multidimensional filter bank theory," *Technical Report*, Caltech, 1992.

[22] N. S. Jayant and P. Noll, *Digital coding of waveforms*, Englewood Cliffs, NJ: Prentice Hall, 1984.

[23] A. Soman, and P. P. Vaidyanathan, "Coding gain in paraunitary analysis/synthesis systems," *IEEE Trans. on Signal Processing*, to be published, May 1993.

[24] I. Djokovic and P. P. Vaidyanathan, "Biorthonormal filter banks: Some necessary conditions and orthonormalization," *IEEE Int. Symp. on Circuits and Systems*, Chicago, May 1993.

[25] P. P. Vaidyanathan and P. -H. Hoang, "Lattice structures for optimal design and robust implementation of two-channel perfect-reconstruction QMF banks," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, pp. 81-94, 1988.

[26] P. -H. Hoang and P. P. Vaidyanathan, "Nonuniform multirate filter banks: theory and design," *Proc. IEEE Int. Symp. on Circuits and Systems*, pp. 371-374, Portland, Oregon, May 1989.

[27] P. P. Vaidyanathan, and S. K. Mitra, "Polyphase networks, block digital filtering, LPTV systems, and alias-free QMF banks: a unified approach based on pseudocirculants," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. ASSP-36, pp. 381-391, March 1988.

[28] H. D'Angelo, *Linear time-varying system: analysis and synthesis*, Allyn and Bacon, Boston, 1970.

List of figures

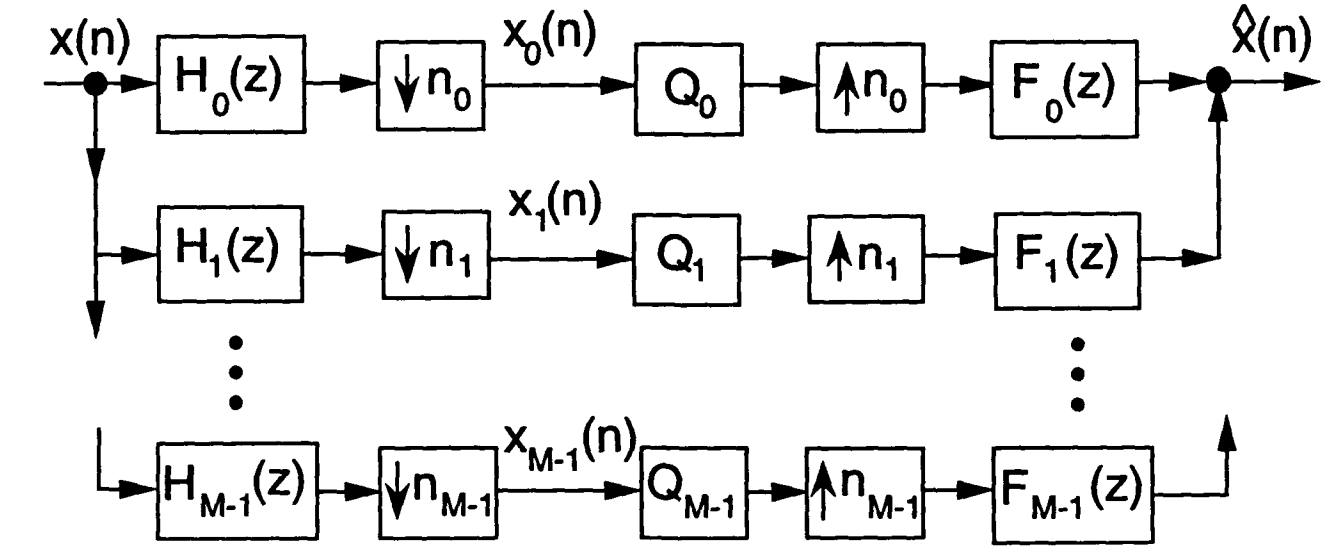
List of tables

FOOTNOTES

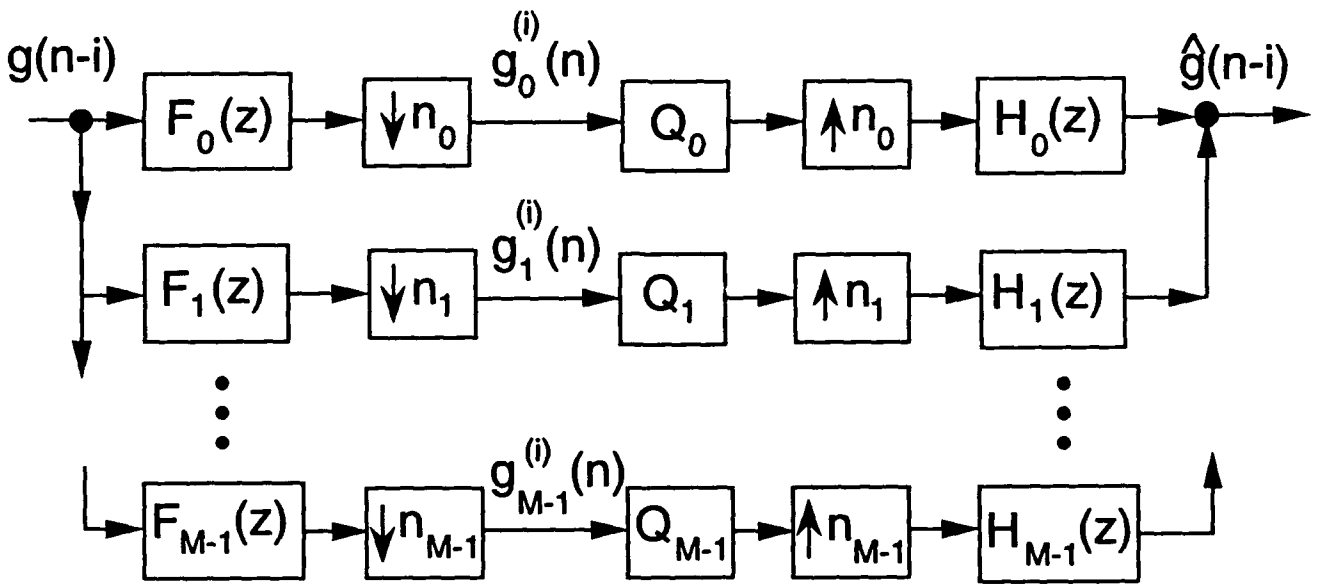
1. Manuscript Received

2. This work was supported by the Office of Naval Research Grant N00014-93-1-0231, and funds from Tektronix, Inc.

3. The authors are with the Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125.



(a)



(b)

Fig. 1.1. The maximally decimated filter bank,
(a) with input x(n),
(b) with input g(n-i).

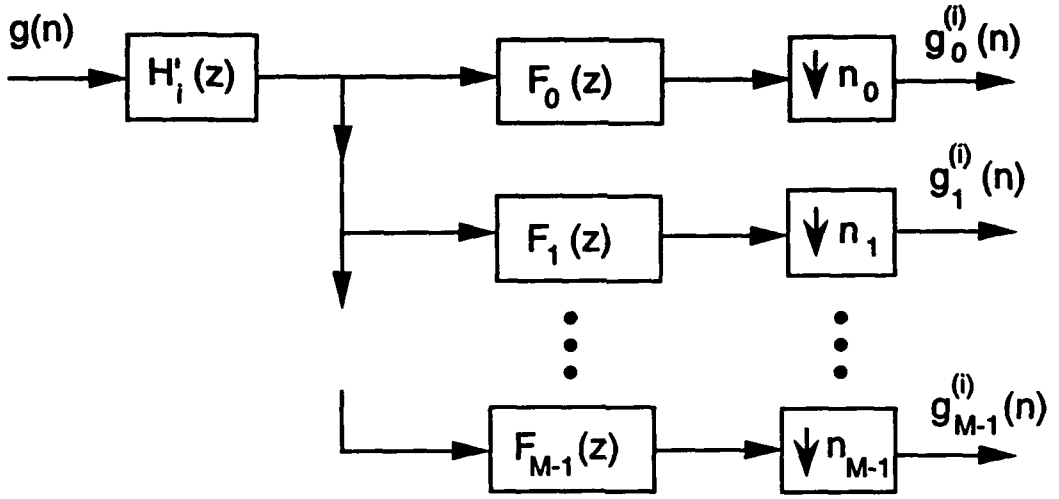


Fig. 2.1. The subband signals of g(n) in two-level FB convolver

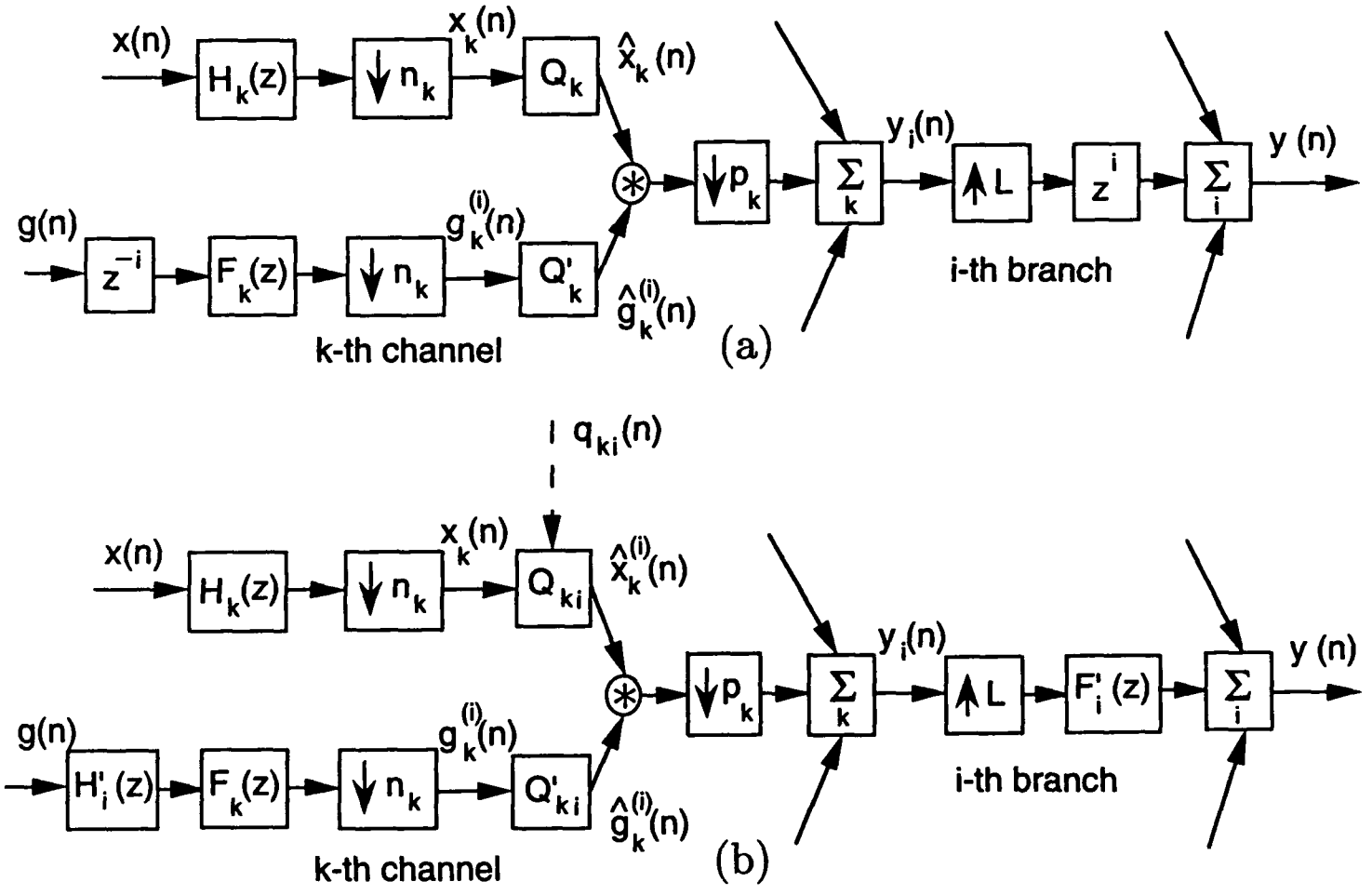


Fig. 2.2. The implementation of filter bank convolvers,
(a) one-level filter bank convolver,
(b) two-level filter bank convolver.
Asterisk * denotes convolution.

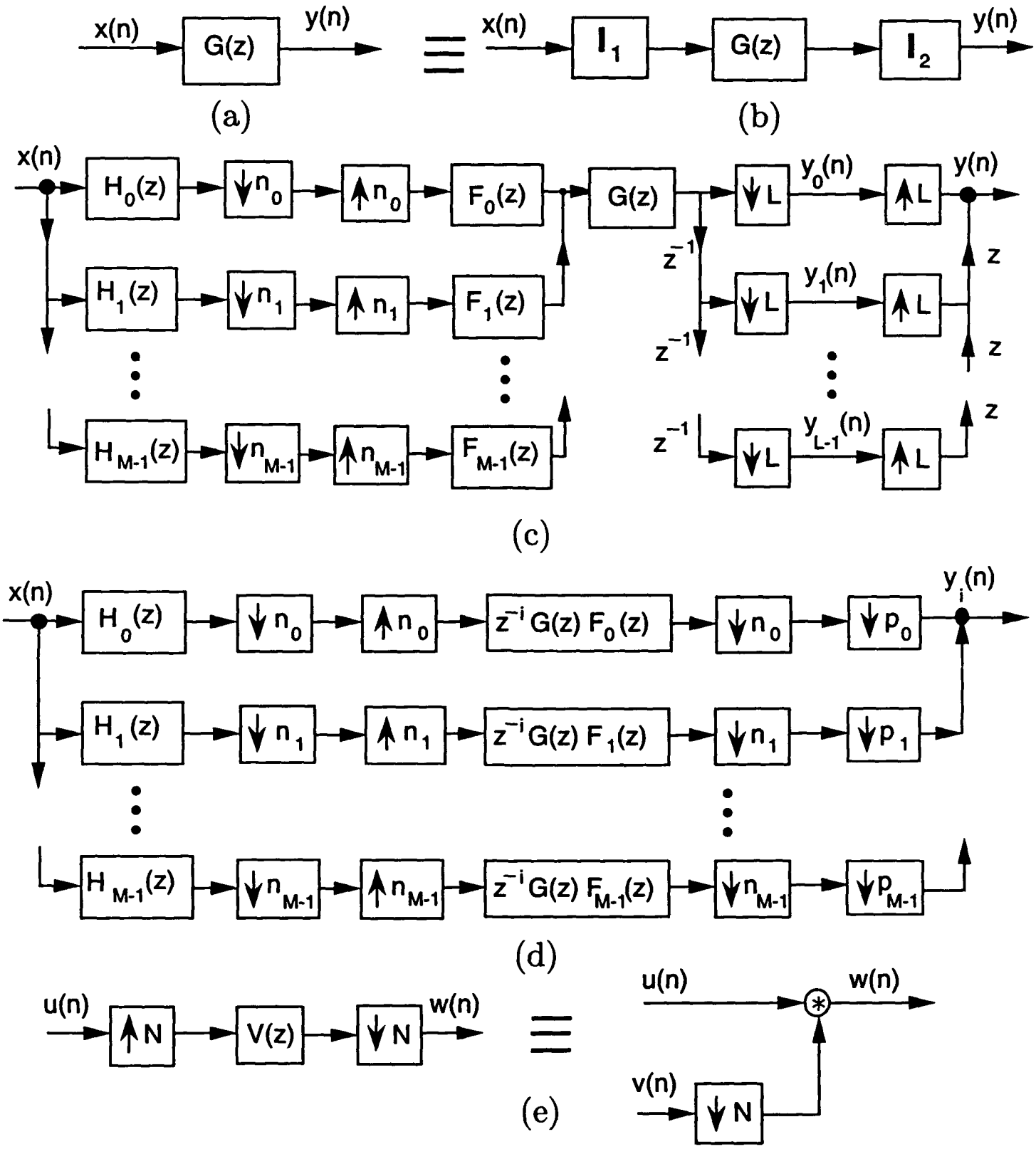


Fig. 2.3. A pictorial proof of Theorem 2.1,
(a) the convolution of x(n) and g(n),
(b) two identity systems inserted,
(c) the identity systems chosen to be two multirate systems,
(d) the i-th branch of (c),
(e) an identity.

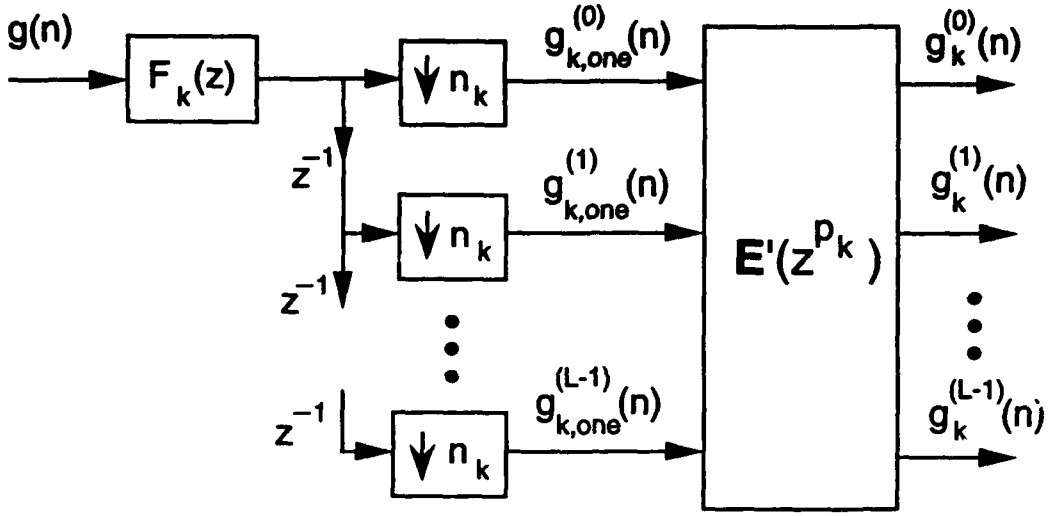


Fig. 3.1. Relationship between the subband signals
of the one-level and two-level FB convolvers.

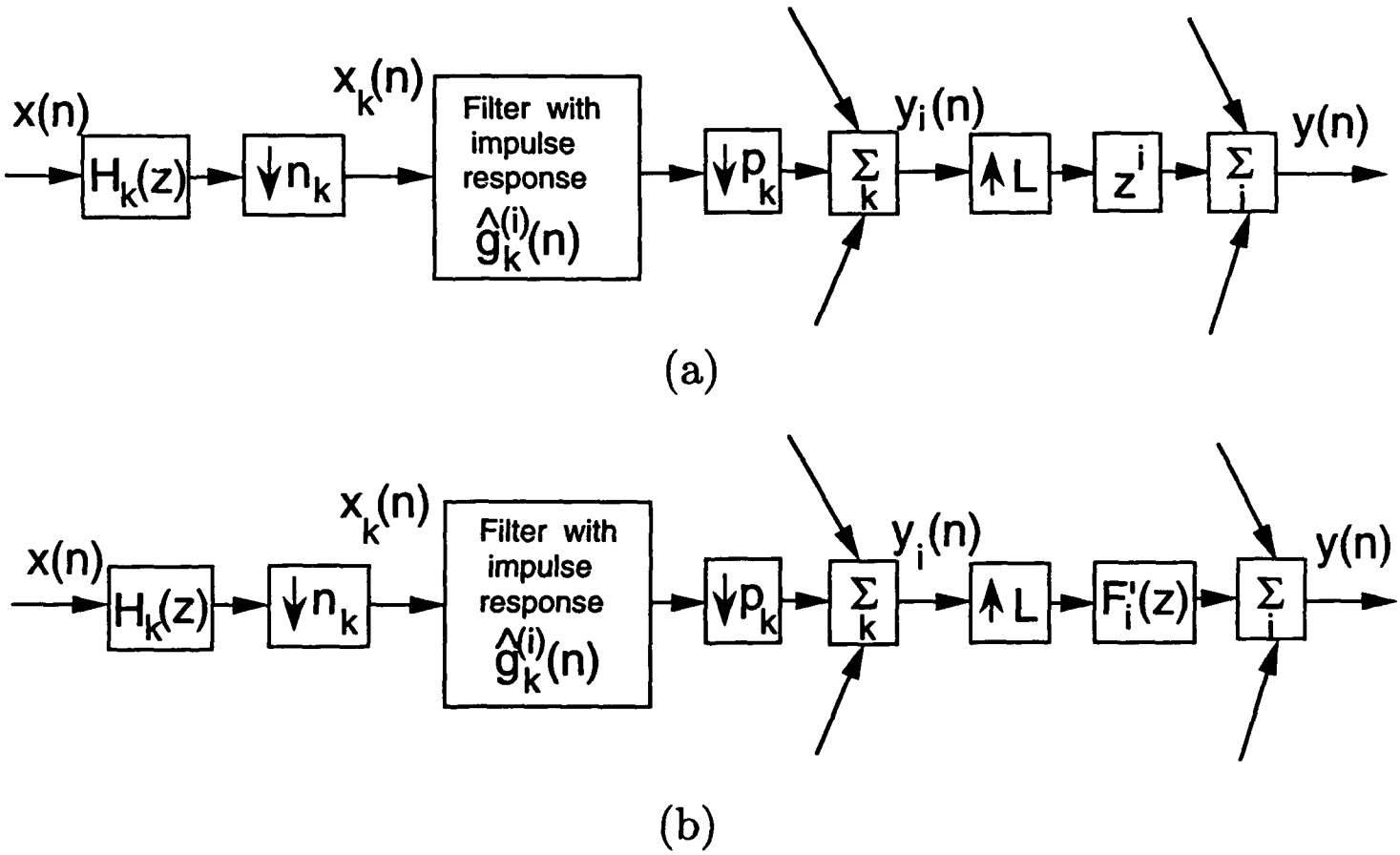


Fig. 4.1. The low sensitivity structures for FIR filters,
(a) with one-level filter bank convolver,
(b) with two-level filter bank convolver.

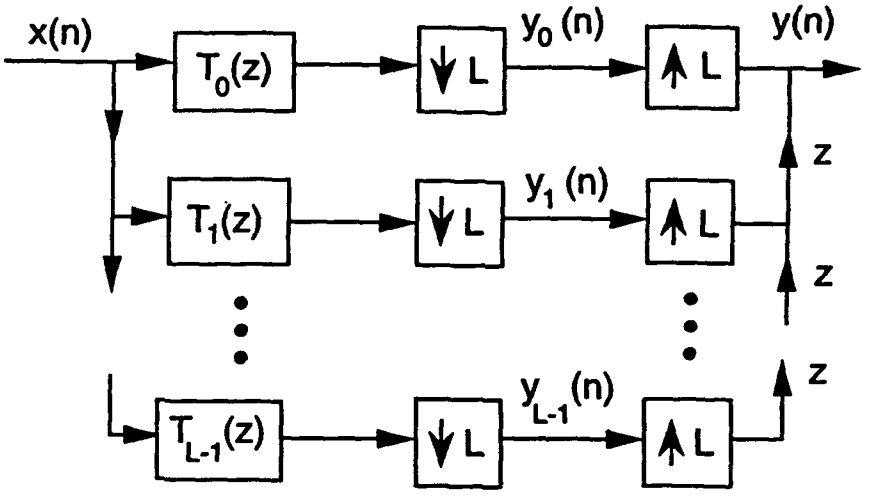


Fig. 5.1. A representation of an LPTV system.

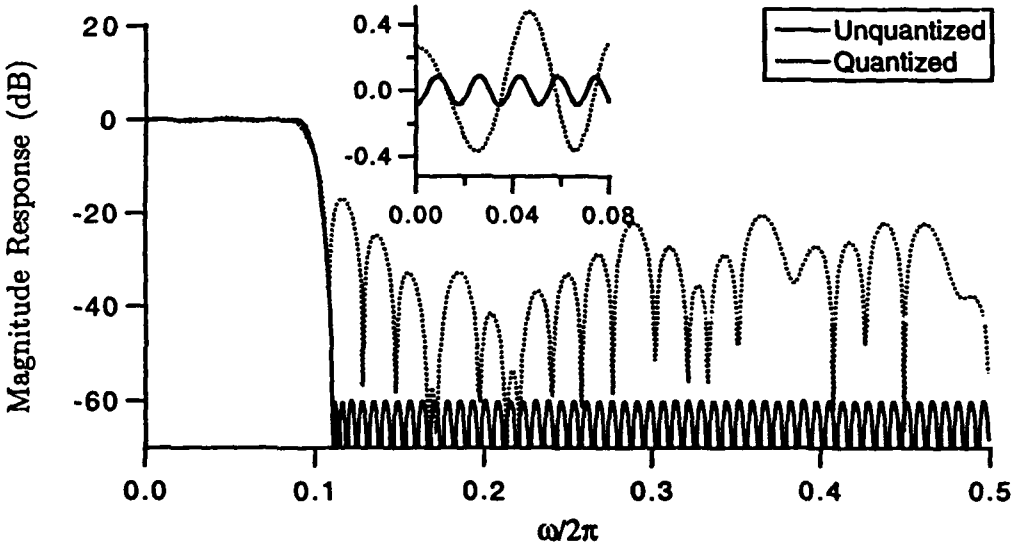


Fig. 5.2. Magnitude response of g(n) with direct quantization to 4 bits, and without quantization.

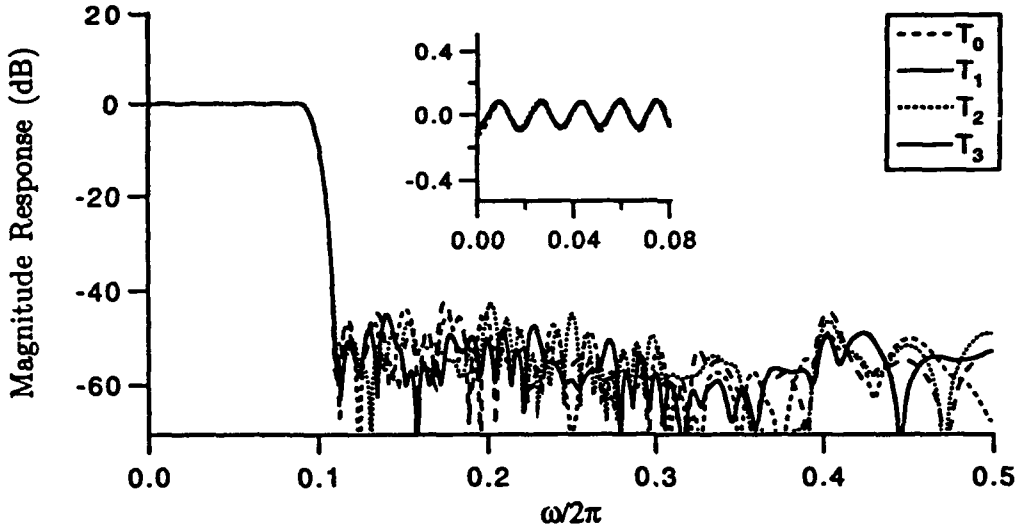


Fig. 5.3. Example 1. Magnitude response of g(n), with subband quantization to 4 bits by using one-level FB convolver.

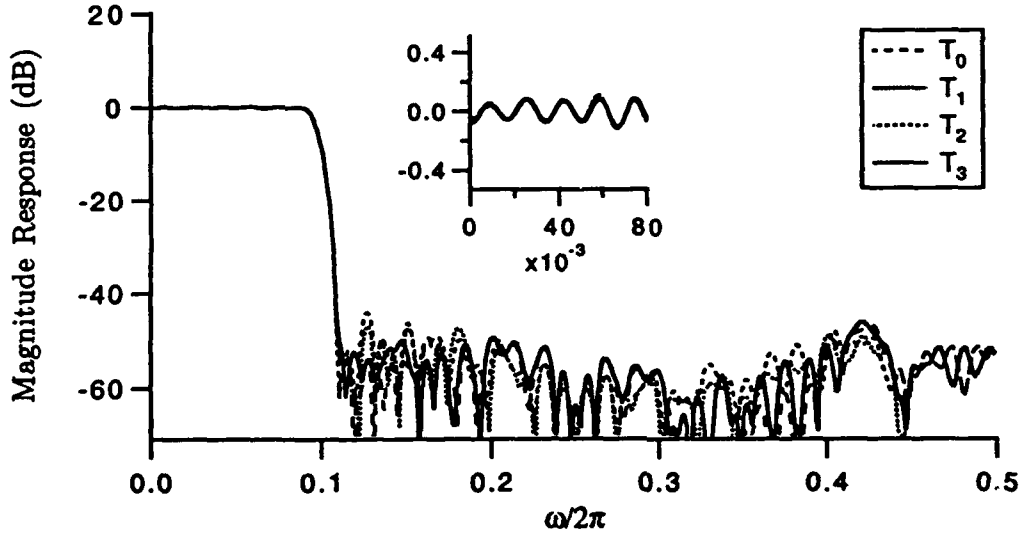


Fig. 5.4. Example 2. Magnitude response of g(n), with subband quantization to 2 bits by using two-level FB convolver.

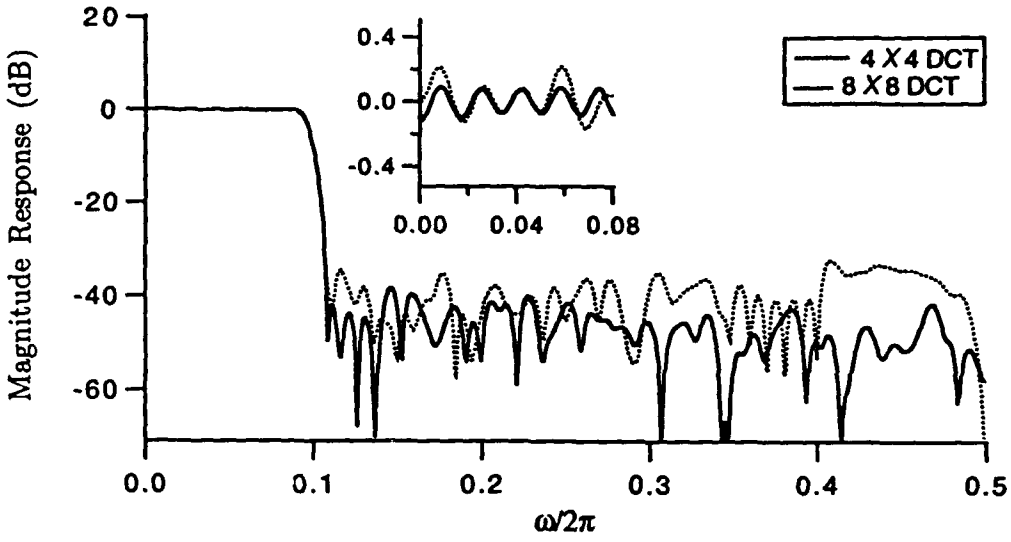


Fig. 5.5. Example 3. Magnitude response of g(n), with subband quantization to 4 bits by using one-level FB convolver.

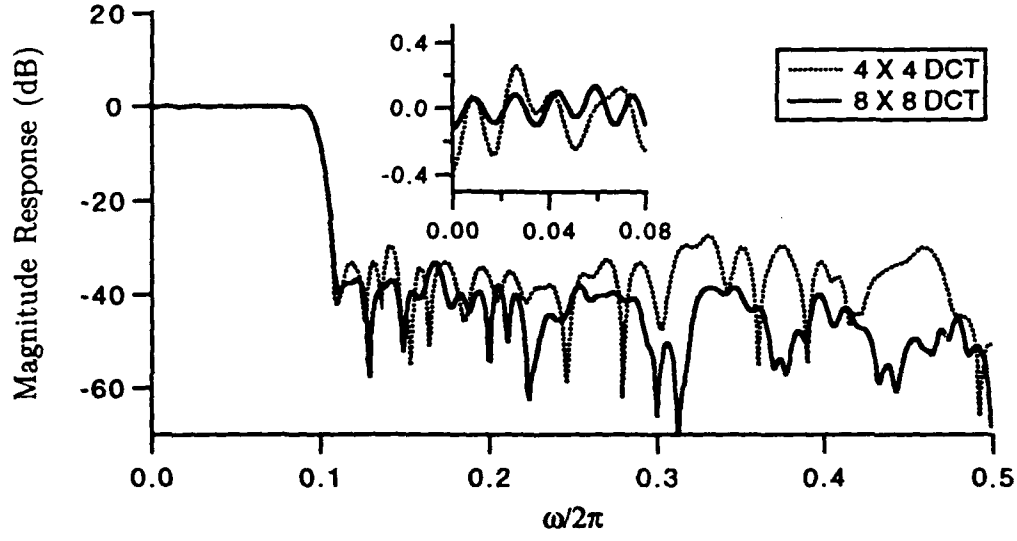


Fig. 5.6. Example 4. Magnitude response of g(n), with subband quantization to 2 bits by using two-level FB convolver.

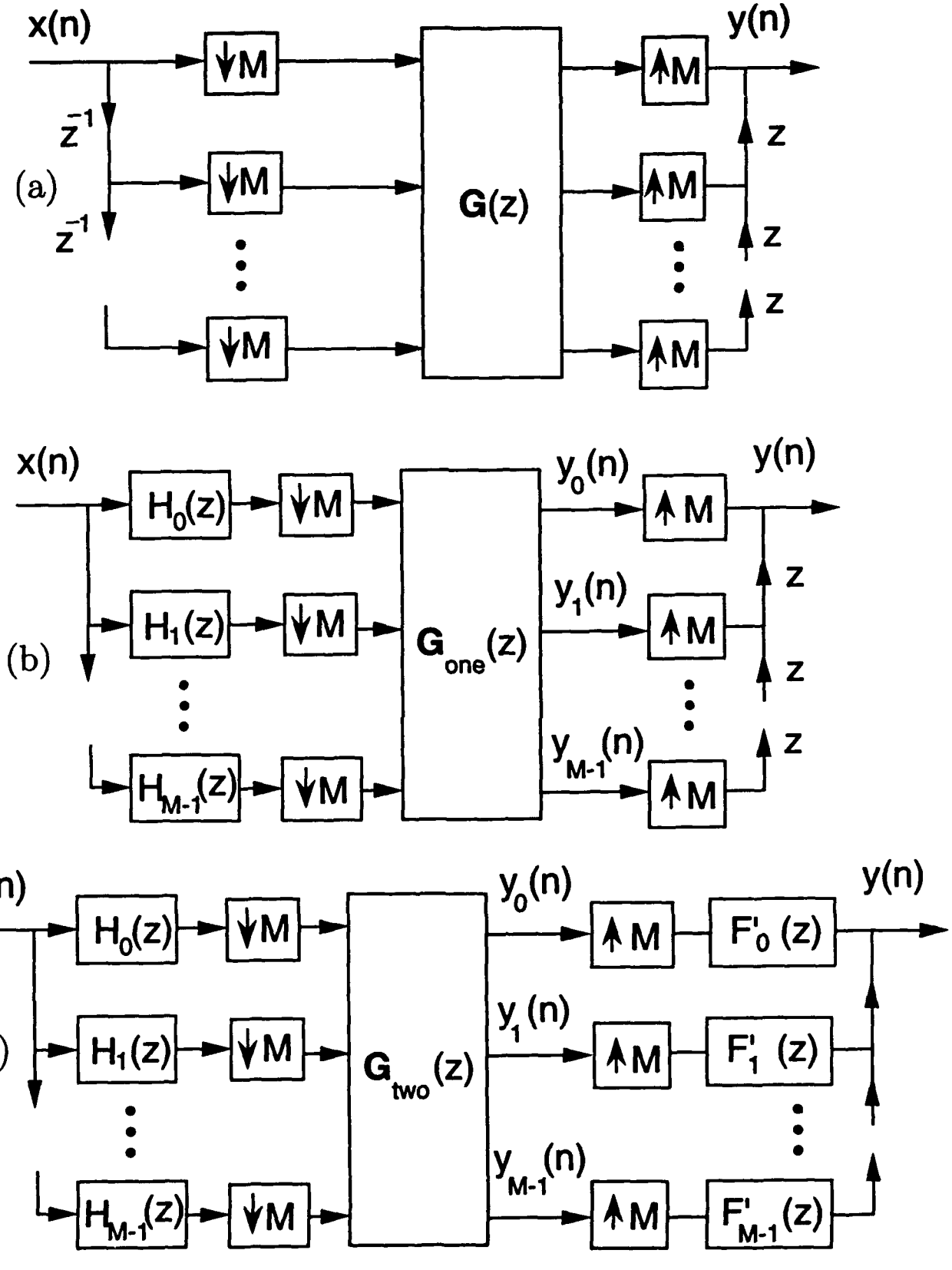


Fig. 6.1. Unified view of block filtering and filter bank convolver,
(a) conventional block filtering,
(b) one-level FB convolver,
(c) two-level FB convolver.

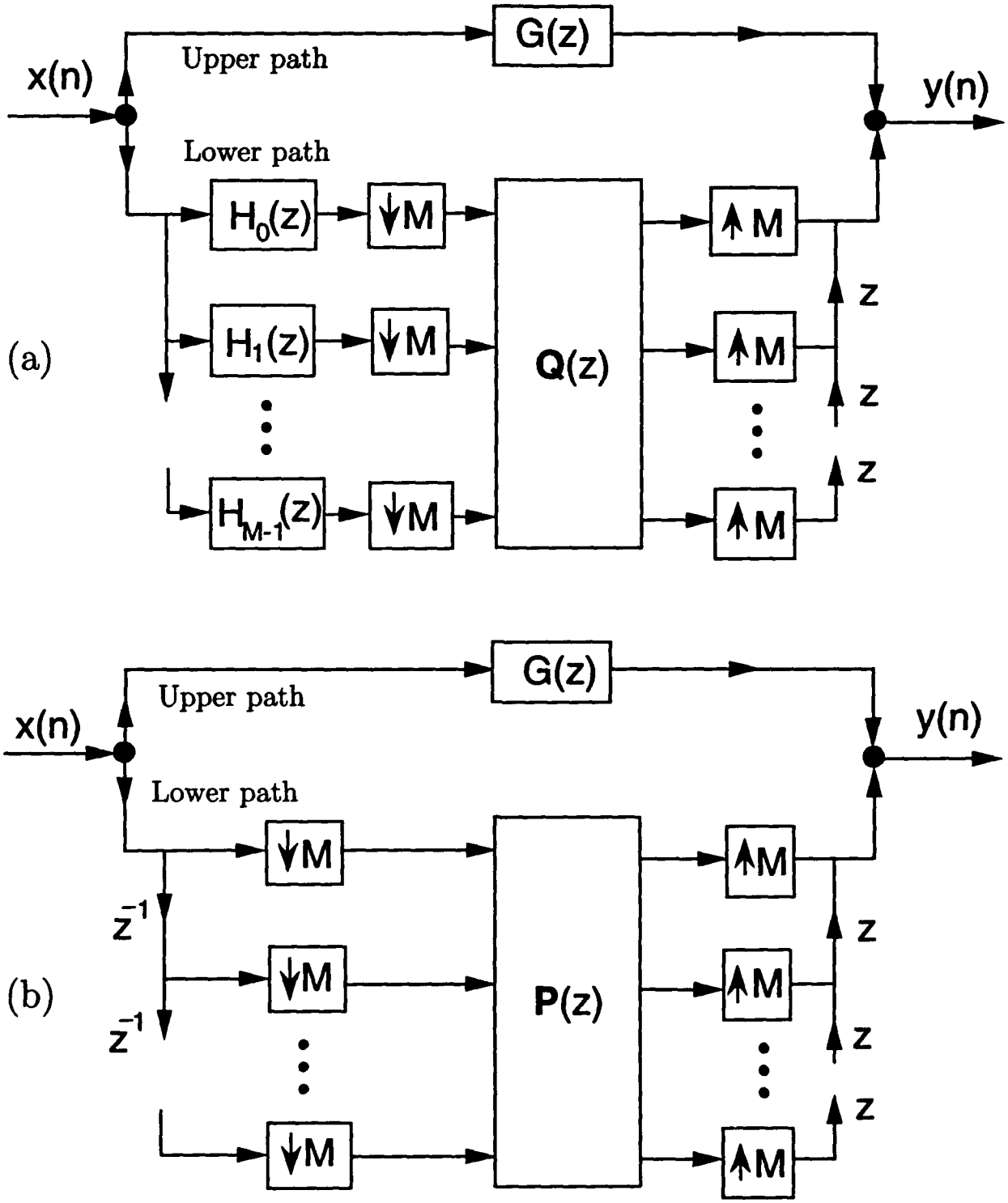


Fig. 6.2. (a) An equivalent representation of Fig. 4.1(a),
(b) a block filter representation.

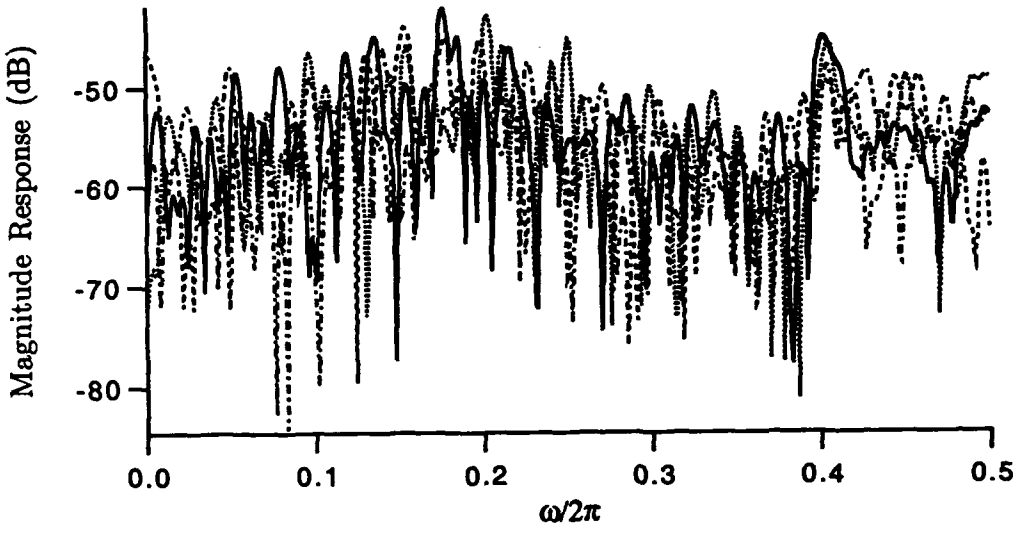


Fig. 6.3. Magnitude responses of the aliasing components.

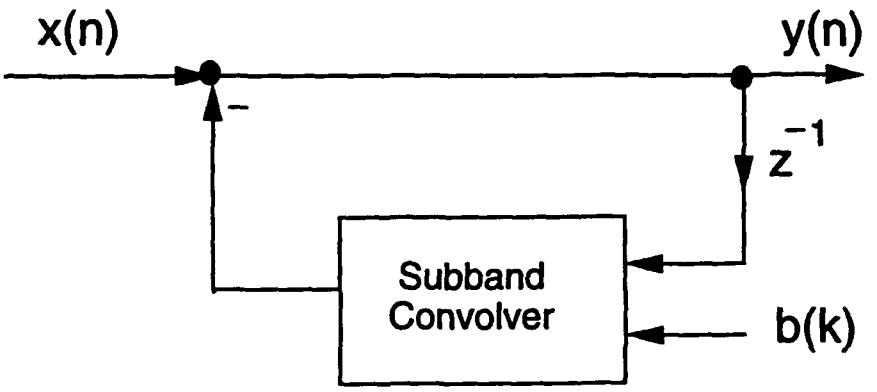


Fig. 7.1. A low sensitivity structure
for an all-pole filter.

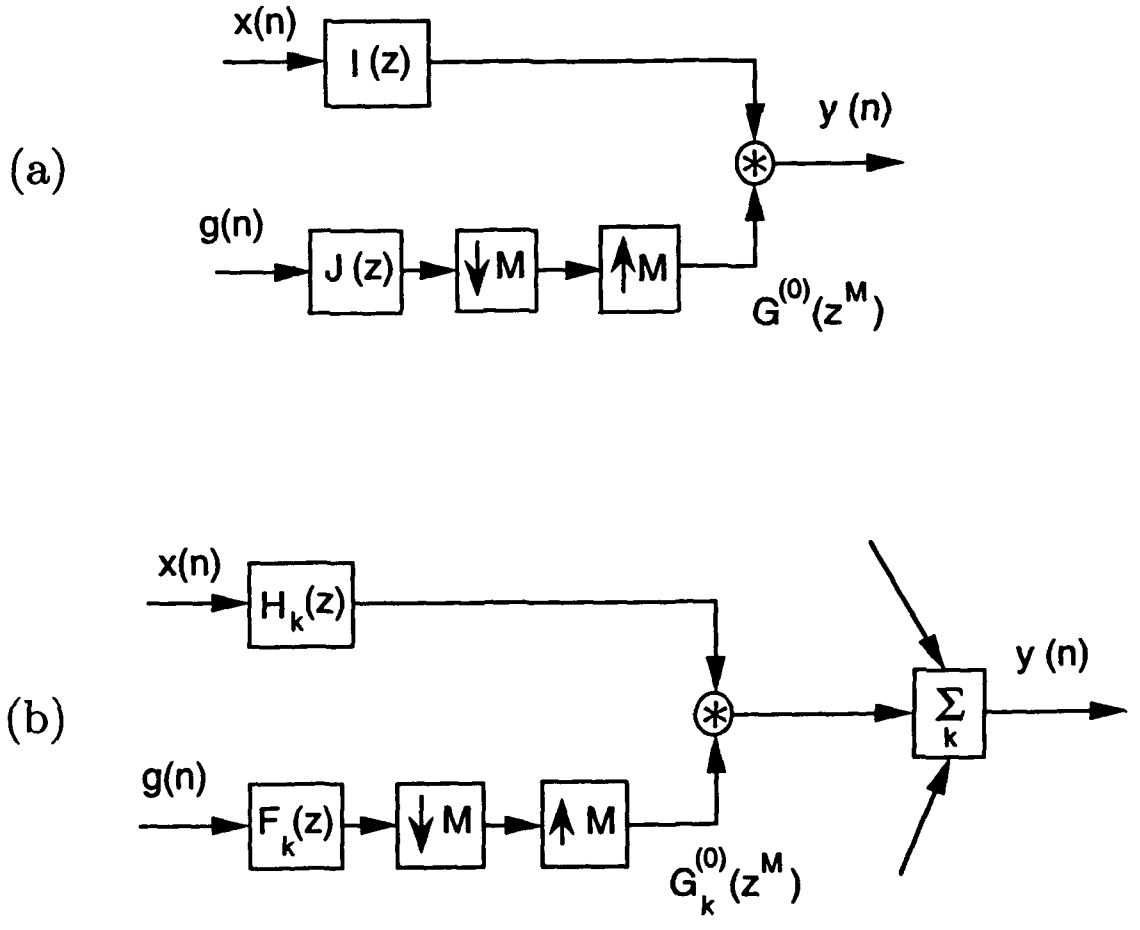


Fig. 8.1. Relationship between convolver and IFIR filter,
(a) implementation of IFIR filter,
(b) implementation of convolver.

	b	δ_s (dB)	δ_p
No Quantization	–	−60	0.010
Direct Quantization	4	−17	0.049
4ch PU Bank: 1-level	4	−42	0.013
4ch PU Bank: 2-level	2	−44	0.015
4×4 DCT: 1-level	4	−32	0.022
8×8 DCT: 1-level	4	−38	0.012
4×4 DCT: 2-level	2	−27	0.035
8×8 DCT: 2-level	2	−33	0.017

Table 5.1. Summary of Examples 1 – 4.

b is the average bit rate.

δ_s and δ_p are the stopband attenuation

and the passband ripple size respectively.

$k=$	0	1	2	3
$i=0$	11	7	2	0
$i=1$	6	4	0	0
$i=2$	2	0	0	0
$i=3$	0	0	0	0

Table 5.2. Example 2. The number of bits b_{ki} allocated to Q'_{ki}

$k=$	0	1	2	3	4	5	6	7
4×4 DCT	7	6	3	0	–	–	–	–
8×8 DCT	9	9	6	3	3	1	1	0

Table 5.3. Example 3. The number of bits b_k allocated to Q'_k

$k=$	0	1	2	3
$i=0$	7	6	3	0
$i=1$	5	5	2	0
$i=2$	2	2	0	0
$i=3$	0	0	0	0

Table 5.4. Example 4(i). The number of bits b_{ki} allocated to Q'_{ki}

$k=$	0	1	2	3	4	5	6	7
$i=0$	9	7	8	5	5	2	2	0
$i=1$	7	9	7	4	4	2	1	0
$i=2$	7	6	6	3	3	1	0	0
$i=3$	3	4	3	1	0	0	0	0
$i=4$	4	3	3	0	0	0	0	0
$i=5$	2	1	1	0	0	0	0	0
$i=6$	2	1	2	0	0	0	0	0
$i=7$	0	0	0	0	0	0	0	0

Table 5.5. Example 4(ii). The number of bits b_{ki} allocated to Q'_{ki}

Filter No.	1	2	3	4	5
$G_{f,\text{one}}$ (dB)	33.2	19.3	17.6	26.4	36.6
$G_{f,expt\ \text{one}}$ (dB)	33.5	19.8	15.5	20.9	34.5
$G_{f,\text{two}}$ (dB)	47.6	28.0	26.5	38.7	54.3
$G_{f,expt\ \text{two}}$ (dB)	49.5	28.5	25.8	36.7	51.9
R_g (dB)	14.4	8.7	8.9	12.3	17.7
$R_{g,expt}$ (dB)	16.0	8.7	10.3	15.8	17.4

Table 5.6. Example 5 and 6. Comparison of coding gain.